Athens University of Economics and Business

SCHOOL of BUSINESS ADMINISTRATION

DEPARTMENT of MANAGEMENT SCIENCE AND TECHNOLOGY

MASTER THESIS

By

NIKOLAOS GATOS

Utilization of Alternative Data And Machine Learning Applications For Credit Scoring at the Greek Hospitality Sector

Internship: Tiresias Banking Systems S.A.

Supervisor: Dr. Nikolaos Korfiatis - Associate Professor

This dissertation is submitted for the partial fulfilment of the requirements for the degree of Master of Science of Management Science and Technology

Athens, June 2022

This page is intentionally blank.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΜΕΤΑΠΤΥΧΙΑΚΟ
ΔΙΟΙΚΗΤΙΚΗ ΕΠΙΣΤΗΜΗ & ΤΕΧΝΟΛΟΓΙΑ
MSc IN
MANAGEMENT SCIENCE & TECHNOLOGY

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ ΓΑΤΟΥ

Η Χρήση Εναλλακτικών Δεδομένων και Εφαρμογών Μηχανικής Μάθησης στην Πιστωτική Αξιολόγηση στον Ελληνικό Ξενοδοχειακό Κλάδο

Πρακτική Άσκηση:

Τειρεσίας Α.Ε. Τραπεζικά Συστήματα Πληροφοριών

Επιβλέπων : Δρ. Νικόλαος Κορφιάτης – Αναπληρωτής Καθηγητής Επιχειρηματικής Αναλυτικής (Επισκέπτης)

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Ιούνιος 2022

This page is intentionally blank.

Declaration

I declare responsibly that the specific master thesis for obtaining the master's degree in Management Science and Technology of the Department of Management Science and Technology of the Athens University of Economics and Business has been written by me personally and has not been submitted or approved by anyone else or undergraduate degree in Greece or abroad. This work, having been prepared by me, represents my personal views on the subject. The sources I referred to for the elaboration of this diploma are listed in their entirety, giving complete references to the authors, including the sources that may have been used by the internet .

Gatos Nikolaos

MSc student in Management Science and Technology

Acknowledgements

I warmly thank my supervising professor, Mr. Nikolaos Korfiatis, for his advices in and support to this dissertation.

I would like to thank my work colleagues without whose help for getting some of the most important data, this research would not have been possible and my manager who also helped me with the handling of sensitive data.

Finally, I want to thank my family and friends for their support and their patience during my studies.

Abstract

Credit scoring refers to the models and procedures that lenders use to determine whether or not to approve a credit or a loan. Credit scoring systems measure the risk of consumer's lending but not their credit. To make these judgments, several strategies and models based on statistical models are employed, and these techniques must be capable of making very accurate predictions which are important for both the lender and the borrower as they determine the equilibrium of money cost. One of the purposes of the credit scoring systems is to forecast the value of a binary variable that indicates whether a consumer will fail to pay back the loan he received or not. Until recently, the most used approach to make such a prediction, was Logistic Regression, however in recent years, Machine Learning methods have been employed to improve the accuracy of the predictions and calculate the Probability to Default (PD). The old-fashioned model also works mainly with financial data that are being collected from the financial institutions – lenders from previous loans and thus it is based on the historical financial behavior of the borrower. The problem that this system is causing, is that borrowers that do not have any financial loan precedent, are very difficult to be scored by the traditional credit scoring models.

The scope of this master thesis is to search, concentrate, prepare, produce and process a dataset of alternative data, by the meaning of not strictly defined financial data of businesses that operate at the hotel -hospitality sector in Greece and then use this dataset in order to test it using Machine Learning models with the goal to develop knowledge for a better credit scoring methods than the traditional statistical technique of Logistic Regression. The outcomes will be used from the Credit Bureau, Tiresias Banking Systems S.A. in order to enhance the databases of the company with alternative data on the specific sector and the knowledge of the Machine Learning models will set comparison measures to the existing traditional models.

For the creation of the dataset that it was later analyzed, there have been used web scrapping techniques, at the site of Hellenic Chamber of Hotels. From this site there were collected qualitative and quantitative data, which later were used as alternative data, by the meaning of non-strictly connected financial data, to train some Machine Learning models.

Keywords: Credit Scoring, Alternative Data, Secondary Data, Hospitality Sector, Hotels, Machine Learning

Περίληψη

Η πιστοληπτική αξιολόγηση αναφέρεται στα μοντέλα και τις διαδικασίες που χρησιμοποιεί ένας δανειστής για να καθορίσει εάν θα εγκρίνουν ή όχι μια πίστωση ή ένα δάνειο. Τα συστήματα βαθμολόγησης πιστοληπτικής ικανότητας μετρούν τον κίνδυνο δανεισμού των καταναλωτών αλλά όχι την πίστωσή τους. Για να γίνουν αυτές οι αξιολογήσεις, χρησιμοποιούνται διάφορες στρατηγικές και μοντέλα που βασίζονται σε στατιστικά μοντέλα, και αυτές οι τεχνικές πρέπει να είναι ικανές να κάνουν πολύ ακριβείς προβλέψεις που είναι σημαντικές τόσο για τον δανειστή όσο και για τον δανειολήπτη καθώς καθορίζουν την ισορροπία του κόστους χρήματος. Ένας από τους σκοπούς των συστημάτων βαθμολόγησης πιστοληπτικής ικανότητας είναι η πρόβλεψη της τιμής μιας δυαδικής αναγνώρισης που υποδεικνύει εάν ένας καταναλωτής θα αποτύχει να αποπληρώσει το δάνειο που έλαβε ή όχι. Μέχρι πρόσφατα, η πιο χρησιμοποιούμενη προσέγγιση για την πραγματοποίηση μιας τέτοιας πρόβλεψης ήταν η λογιστική παλινδρόμηση, ωστόσο τα τελευταία χρόνια έχουν χρησιμοποιηθεί μέθοδοι Μηχανικής Μάθησης για τη βελτίωση της ακρίβειας των προβλέψεων και τον υπολογισμό της πιθανότητας αθέτησης (PD). Το παραδοσιακό μοντέλο λειτουργεί επίσης κυρίως με οικονομικά στοιχεία που συλλέγονται από τα χρηματοπιστωτικά ιδρύματα – δανειστές από προηγούμενα δάνεια και επομένως βασίζεται στην ιστορική οικονομική συμπεριφορά του δανειολήπτη. Το πρόβλημα που προκαλεί αυτό το σύστημα είναι ότι οι δανειολήπτες που δεν έχουν προηγούμενο χρηματοοικονομικών δανείων είναι πολύ δύσκολο να βαθμολογηθούν από τα παραδοσιακά μοντέλα πιστωτικής βαθμολόγησης.

Σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η αναζήτηση, συγκέντρωση, προετοιμασία, παραγωγή και επεξεργασία ενός συνόλου δεδομένων εναλλακτικών δεδομένων, με την έννοια των μη αυστηρά καθορισμένων οικονομικών δεδομένων των επιχειρήσεων που δραστηριοποιούνται στον τομέα των ξενοδοχείων - φιλοξενίας στην Ελλάδα και στη συνέχεια να χρησιμοποιηθεί αυτό το σύνολο δεδομένων προκειμένου να τεσταριστεί χρησιμοποιώντας μοντέλα μηχανικής μάθησης με στόχο την ανάπτυξη γνώσεων για καλύτερες μεθόδους βαθμολόγησης πιστοληπτικής ικανότητας από αυτήν της παραδοσιακή στατιστική τεχνικής της λογιστικής παλινδρόμησης. Τα αποτελέσματα θα χρησιμοποιηθούν από την εταιρεία πιστωτικών αξιολογήσεων «Τραπεζικά Συστήματα Πληροφοριών Τειρεσίας Α.Ε.» . προκειμένου να ενισχυθούν οι βάσεις δεδομένων της εταιρείας με εναλλακτικά δεδομένα για τον συγκεκριμένο κλάδο και η γνώση των μοντέλων μηχανικής μάθησης να θέσει νέα μέτρα σύγκρισης για τα υπάρχοντα παραδοσιακά μοντέλα.

Για τη δημιουργία του συνόλου δεδομένων που αναλύθηκε αργότερα, χρησιμοποιήθηκαν τεχνικές διάλυσης ιστού, στον χώρο του Ξενοδοχειακού Επιμελητηρίου Ελλάδος. Από αυτόν τον ιστότοπο συλλέχθηκαν ποιοτικά και ποσοτικά δεδομένα, τα οποία αργότερα χρησιμοποιήθηκαν ως εναλλακτικά δεδομένα, με την έννοια των μη αυστηρά συνδεδεμένων οικονομικών δεδομένων, για την εκπαίδευση ορισμένων μοντέλων μηχανικής μάθησης.

Λέξεις Κλειδιά: Πιστωτική αξιολόγηση, Εναλλακτικά δεδομένα, Κλάδος Φιλοξενίας, Ξενοδοχεία, Μοντέλα Μηχανικής Εκμάθησης

Contents

Index of Images

Index of Tables

1. Introduction

1.1 Credit Scoring

"*The future will be like the past*", is a common phrase that governs most human decision-making, however our uncertain environment contradicts this premise, with little to major variations. Predictive models are powerful, but they're not flawless, especially when essential data is lacking from the analysis—for example, the economy, competition, and legislation. Additionally, using a model may alter the behavior it is predicting, such as fraud, hastening its invalidation. Regardless, the estimations that follow are valuable decision-making tools.

Credit scoring can be defined as a collection of decision models and underlying procedures that help lenders provide consumer credit. In other words, it's a form of risk-modelling used to provide ratings used in credit intelligence, and to a not insignificant extent, in mass financial-surveillance. These methods determine who can get credit, how much credit they will get, and what operational tactics will improve the borrowers' creditworthiness to lenders. Apart from its original purpose, Credit scoring tools have gone far beyond their primary goal of determining credit risk. Evaluating the risk-adjusted profitability of account relationships, determining the initial and continuing credit limits accessible to borrowers, and aiding in a variety of loan servicing operations, such as fraud detection, delinquency intervention, and loss reduction. (Anderson, 2022)

Credit scoring models offer some important advantages that are including the reduction of credit analysis and credit assignment expenses through an affective and rapid decision making process, greater likelihood of credit repayment and less potential risk (Koutanaei et al., 2015). Changes in technology have increased the depth and breadth of available data, enabling the use of previously inviable predictive techniques. The most used technique for Credit scoring models is Logistic Regression. But in this involving field, new models are being developed with Machine Learning (ML) and Artificial Intelligence (AI) methods, which combine statistics and computer science. Traditional statistics are 'old school,' requiring a lot of work to derive insights from limited data while but with a strong theoretical foundation. Computer science is 'new school'; it uses brute-force computing to process both large and small amounts of data, combining classic statistics with newly developed methodologies. Machine Learning focuses on predictions and their practical application, whereas statistics focused on study and comprehension.

Credit scoring has found widespread usage in a variety of sectors, including statistical techniques for prediction and classification problems. A number of processes must be incorporated in corporate credit scoring models, ranging from collecting and preparing pertinent data to predicting a credit score using a formula induction method, as well as constructing, monitoring, and recalibrating the scorecard. For instance, data collection and preparation for missing value management and the selection of a predictive set of explanatory factors. Once a data collection has been prepared, a number of prediction methods can be used to estimate various components of credit risk.

1.2 Dissertation Objectives

Tourism is one of the largest sectors of the Greek economy and it is considered to be the "heavy industry" of the country. This is due to the 20,8% of the participation of the sector to the GDP, by the 946.200 workers that service around 31,3 millions of tourists (cruise tourists are excluded from this data) according to Association of Greek Tourism Enterprises (*Βασικά Μεγέθη του Ελληνικού Τουρισμού 2019*). The summer period of 2022, after two "lost" touristic seasons due to the global pandemic of Covid-19, it is being expected that the previous numbers will exceed the levels of 2019 and make the sector flourishing again against the geopolitical factors that the continent faces.

In this environment and due to the inexpensive money- credit that is been offered by the Recovery and Resilience Facility Loans of the EU, a lot of hotel businesses in Greece are asking Tiresias S.A. Banking Systems for their Credit Assessment in order to complete their business plan and ask for investment loan from the Greek banks. The problem is that a lot of these businesses, cannot provide sufficient financial data as for the traditional credit scoring models to produce extremely reliable credit scores. The main reason for this phenomenon is that a lot of this hotel businesses are newly established, or they have not got any previous credit history.

The lack of financial data that are being used for the traditional credit scoring assessments, is one of the two problems that this master thesis is trying to deliberate. The second problem has to do with the model that is being used and the traditional technique of the logistic regression, that cannot produce reliable scores for all these businesses.

This thesis tries to respond to the question of whether it is possible to produce trustworthy credit scores, using alternative data, by the mean of non-financial data combined with some of the most used Machine Learning algorithms. In this research and due to the reasons, that where previous analyzed, we will focus in the use of data from the Touristic Sector and most specifically the Hotels, in order to try to develop new accurate and robust credit scoring models with Machine Learning. The thesis will not focus primarily at the models and their fine tuning, but it is focusing primarily on the collection of the data and tries to answer the question of whether the data can have a good predictive importance for credit scoring models specialized in hotel sector.

1.3 Contribution

One of the biggest problems that credit risk assessment companies have to cope with is the collection of accurate and reliable data that can guide to the developement of highly predictive credit risk models. The R&D team of Tiresias S.A., of which the writer of this dissertation is part of, is constantly trying to develop new models that can increase the added value of the company's products and to maintain its brand name for credit scoring assessments in the highest level as is it

today. The results of this dissertation are filling some parts of this constant adventure of discovering and understanding credit scoring and data processing.

For many years, Tiresias S.A. collects data for individuals and companies, from various sources, such as lawyers at the law courts, balance sheets, loan data from the banks, to name a few. The datasets that are created though, are sometimes very difficult to merge and combine them in order for the data scientists to develop better scoring models. With this research, the idea of geocoded joins between two datasets, was tested and resulted very good results, that led the R&D of the company to apply the acquired knowledge to process data sets for other projects and to operationalize the method for some data tables. The company also acquired knowledge as for the importance of Machine Learning in credit score modeling, which will be used in the research for other projects that are under development.

As for the target audience, they can get familiar with a real life application of data processing and credit score classification. If the results of models were slightly better, the models could be used to assess investments at the hospitality sector that are currently in development in Greece. It is highly possible that the future work proposals would lead to better models that will have active role in decision making for assessing investment portfolios and calculate credit risk using non-financial but alternative data.

By this research

1. There was developed a big data set with qualitative and quantitative characteristics of 10111 hotels in Greece
2. There was created a data set that contains a set of some of the hotels along with their credit scoring from a previous financial based analysis from Tiresias S.A.
3. There were trained 7 Machine Learning classifiers with the last dataset, namely:
    a. Random Forest
    b. Naive Bayes
    c. Decision Tree
    d. Gradient Boosting
    e. K-Nearest Neighbour
    f. Logistic Regression
    g. Support Vector Machine
4. The assessment of the models was conducted with the Area Under the Curve (AUC or AUROC) metric

1.4 Dissertation Structure

To this end, this dissertation is structured as follows:

1. In the first chapter of this dissertation, Credit Scoring is introduced to the reader.
2. At chapter 2 there is a brief discussion of prior literature that has been conducted in this field of Credit Scoring and the applications of Machine Learning to this financial field as well as some historic information and theory of credit scoring. This is also the chapter

where the reader can study most of the specific theory needed in order to build up the knowledge for the following chapters.

3. In chapter 3 there is a detailed description on how the alternative data were collected, how the datasets were cleaned and handled and how multiple difficulties in the creation of the final dataset were overcame.
4. In chapter 4 are presented the Machine Learning model outputs and the relevant metrices of the computational experiments that were conducted. This is also the chapter where the outcomes of the experiments are commented and possible changes as for the data handing or the models are discussed.

2. Theoretical Background and Key Concepts

2.1 Credit Risk Theory

Credit risk is defined as the probable inability or reluctance of borrowers to fulfill their obligations. We refer to 'default' as an event that implies a big rise in that potential. Loans are considered 'performing' before the occurrence (which must be defined), because lenders book the interest as income; after that, income cannot be recognized for 'non-performing' loans. A missed payment (traded debt securities), after a pre-determined time since the scheduled payment date (banking), or an event indicating gloomy clouds on the horizon (e.g. application for liquidation) can all trigger Default. The situation must be managed once it has been triggered to ensure that the problem is resolved or that recoveries are maximized loans. Most of our focus is on default prediction, which is the backbone of most credit rating and decision making, with severity second.

Credit risk, as stated in the glossary, generally derives directly from the borrower/standing, counterparty's but it can be increased or decreased by the type of the transaction.(Donaldson, 1989) A long-term loan offers a higher interest rate. A long-term loan carries more credit risk than a short-term loan, for example, because more businesses fail in ten years than in one; or an unsecured loan carries more credit risk than a secured loan to the same borrower, because secured lenders benefit from their security, whereas unsecured lenders may recover nothing. Credit risk may be so significantly dependent on the transaction with non-recourse financing or loans that rely primarily on collateral that there is no true borrowing entity.

2.2 Basic Terminology Of Credit Scoring

A credit scoring is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of an individual. We could say that now days, credit scoring is an application of mathematical methods that measure Credit Risk. A credit score is primarily based on a credit report, information typically sourced from credit bureaus.

In this section are given some basic definitions of Credit Scoring theory.

- **Credit:** Although we identify it with 'buy now, pay later,' its Latin root—'credo,' means 'trust in' or 'rely on.' We trust people to honour their obligations.
- **Risk:** Is the exposure to danger, severity, or expected added value, harm, loss, or missed opportunities. Unless they are gambling addicts or extreme adrenaline junkies, most people want to either reduce risk or find the best risk/reward trade-off.
- **Credit Bureau:** An agency that provides a service, usually involving knowledge or information.

- **Rating:** A single label or number that summarizes information allows subjects to be sorted/ranked according to some perceived quality or real performance.
- **Score:** A rating presented as a number, possibly with as many as 999 possible values, whether derived via the assignment and totaling of points or some other means.

Our focus is on automated assessments that make judgments based on estimates or ratings. The bigger the implications of decisions, the larger the competencies required. Credit reports are a good place to start, but they can be confusing. Credit ratings are ranking tools, no matter how well-intentioned, that rank-order subjects based on whether or not we will get our money back, and/or how much we will get back—with the word 'whether' being strongly tied to probabilities or odds. Empirical models, human judgment, or any mix of the two are used to award such evaluations.(Orgler, 1970)

2.3 Some History Of Credit Scoring Industry And The Modern Market

Credit scoring is an industry that saw its highest development the last 50 years, with the development of the technology of highly intelligent computational system that permit the analysis of big data sets. Another import factor for the thrive of the industry, was the flourishing global economy, after the WW2 and the stable economic environment that is was formed after the great war. That been said, the development of the sector is summed up by the following chronological events in the following table.

For the next table: "CC" is the country code, and "W" is whether it is: m) mercantile/trade credit; c) consumer credit i) bond investments; A) association of credit bureaux; G) guild, or professional society. A bold font is used to highlight key names."

Table 1: Credit scoring historic events. (Anderson, 2022)

CC	W	Year	Event
UK	c	1776	London Society for the Protection of Trade Against Swindlers & Sharpers, founded
UK	c	1803	Mutual Communication Society of London, founded by several tailors
UK	m	1827	Manchester Guardian Society founded
US	m	1841	Mercantile Agency founded by Lewis Tappan
UK	c	1842	London Association for the Protection of Trade (LAPT) founded for West-end carriage trade
US	m	1849	John M. Bradstreet & Sons founded in Cincinnati
US	m	1859	R.G. Dun purchases the Mercantile Agency

US	b	1862	Poor's Publishing Co. founded by Henry Varnum Poor
UK	A	1864	National Association of Trade Protection Societies formed in England
US	c	1869	Retailers Commercial Agency (RCA) founded in Brooklyn NY
DE	m	1882	Verein Creditreform zum Schutze gegen schädliches Creditgeben founded
US	m	1888	Credit Clearing House founded
US	G	1896	National Association of Credit Men founded
US	c	1897	Chilton Corp. founded in Dallas TX by James Chilton
US	c	1899	Retail Credit Co. (RCC) founded in Atlanta GA by Cator and Guy Woolford
ZA	m	1901	R.G. Dun establishes Cape Town office, which becomes ITC in 1986
US	A	1906	National Association of Credit Bureau founded
US	I	1909	John M. Moody does first bond ratings; inc. as Moody's Investor Services in '14
US	I	1913	John Knowles Fitch establishes Fitch Publishing, today Fitch IBCA
DE	c	1927	Schufa Holding AG formed in Germany by a group of banks and retailers
US	c	1932	Michigan Merchants Co. founded, and later renamed Credit Data Corp. (CDC)
US	m	1933	Dun & Bradstreet merger
DE	r	1934	Evidenzzentrale für Millionenkredite founded, first public credit registry
US	I	1941	Standard & Poor's created from merger of Poor's with Standard Statistics
UK	c	1965	LAPT renamed to United Association for Protection of Trade (UAPT)
US	c	1968	TRW buys CDC, to establish TRW Credit Data (TRW-CD)
US	c	1968	TransUnion founded by Union Tank Car Company (UTCC)
US	c	1975	RCC renames itself to Equifax
UK	c	1980	Consumer Credit Nottingham (CCN) founded by Great Universal Stores (GUS)
IT	c	1988	Centrale Rischi Finanziari (CRIF) founded in Bologna, Italy
UK	c	1994	Equifax buys UAPT-Infolink
UK	c	1996	CCN and TRW-IS&S merge to become Experian

IS	c	1997	Lánstraust ehf founded in Reykjavik, later renamed Creditinfo
UK	c	2000	Callcredit founded in Leeds, initially focused on marketing information.

The most famous and widely used method, for retail market credit score segment, is FICO scores, a credit score report generated by Fair, Isaac & Co. (short: FICO), a data analytics business located in San Jose, California, that launched the first commercially successful application scorecards in the early 1960s. FICO was founded in 1956 engineer William Rodden Fair (1922–96) and mathematician Earl Judson Isaac (1921–83), two ex-employees of the Stanford Research Institute in California. Their first contract in '57 was to develop a billing system for Carte Blanche, a credit card offered by Conrad Hilton's hotel chain. (Anderson, 2022)

They recognized that Linear Programming could be used to create a prediction model and suggested 'credit scoring' to 50 prospective clients via mail at 1958. The only reply was from American Investment Company, one of the largest personal cum instalment financing institutions in the United States, which served manufacturing and blue-collar employees. That same year, FICO developed its first scorecard at its Public Finance Company of Missouri, followed shortly after by a second scoreboard for Louisiana, by 1969, they had distinct scorecards for seven areas. The primary purpose was to expedite the processing of the large amount of loan applications in a thriving economy, with loss reduction as a secondary goal.

Since the early 1990s, the term has been synonymous with those offered by the Big Three credit bureaus (Experian, TransUnion, and Equifax). Their credit ratings have a restricted but deep obligational foundation, which means they are produced using a large amount of data about one area of consumer behavior: how they handle their credit. Americans (among others) have a near obsession with them, which is appropriate, given the importance of consumer credit in their economy.

Regardless matter how restricted the data in credit reports is, there are hundreds, if not thousands, of factors in each credit report, and no human consensus on what is most significant. As a result, they are mined for data in models that forecast delinquency or default, and the likelihood is known as a credit score. FICO, by being the first to develop such a score and cleverly incorporating it into a relative score ranging from 300 to 850, has become a norm in many US marketplaces. The computer-generated score is impervious to explanations of what this was done or happened, or even pointing out flaws in the data. FICO gives hints, but not explanations, on what makes their ratings operate in order to minimize wrath while still selling their proprietary methodology.

The earliest scores were general ('Classic'); they were updated on a monthly basis. Later scores were tailored to I industry (bank card, auto loan, mortgage, installment financing, personal loan); ii) lifecycle (application, collection); or iii) sub-population (XD for thin-file customers, billing data for utilities, cable TV, and mobile phones, SBSS for small companies). Scores can also change from bureau to bureau due to varied data from each bureau's subscriber base, but FICO 8 allegedly includes data from all American bureaux. An American can have over 50 distinct FICO scores at any given moment, depending on which model is used—adding to the consumer's

perplexity. There are several publications available that attempt to advise on how to manage the scores, the most of which is plain sense.

There is minimal or no variety in what goes into the FICO score. There are some information that are provided from FICO. The hints are for the general version, rounded to the closest five percent to avoid frequent updating; real numbers will vary depending on which model version is used and by whom. Labels can also differ based on how an author interprets them. The generally used statistics and descriptors are: 35% —payment history; 30% —amounts owing; 15% —length of credit history; 10% —credit mix; and 10% —enquiries/new credit. (Anderson, 2022)

The other big format for retail market credit score segment in the industry is VantageScore, which is a consumer credit-scoring methodology developed by a collaboration of the three main credit agencies (Equifax, Experian, and TransUnion). VantageScore Solutions, LLC, a separate firm founded in 2006 and jointly owned by the three agencies, manages and maintains the model. VantageScore models compete with Fair Isaac Corp.'s credit scoring formulas (FICO). VantageScore models, like FICO models, function on data held in consumer credit files maintained by the three major credit agencies. VantageScore and FICO models employ statistical research to forecast the possibility of a consumer defaulting on a loan. The risk of loan default is represented by three-digit ratings in the VantageScore and FICO models, with higher scores representing lesser risk. Scores from one system cannot be converted into the other because VantageScore and FICO employ separate, proprietary analytical procedures. VantageScores are widely used by credit card issuers, and secondly by both installment loan and fintech lenders.(Gravier, 2021)

In the segment of corporate modelers, the most well known companies, are (1) JP Morgan—and its RiskMetrics and CreditMetrics; (2) KMV—who used Merton’s model to assess default probabilities; and (3) Moody’s—and its Credit Research Database. CreditMetrics was launched in 1997. It was developed largely by Greg Gupton, based heavily upon transition matrices and price movements as credit ratings change (Gupton et al. 2007). In 1989, KMV was founded from Stephen Benson Kealhofer (Princeton—PhD economics), John Andrew McQuown (Harvard—MBA) and Oldich Alfons Vaek (Charles University, Prague—PhD mathematics). It soon rose to notoriety due to its default probability predictions and later it was purchased by Moody's Investor Services in 2002 for US$210 million, and is now part of Moody's Analytics.

The fourth in the rank, in terms of size, leader company in credit risk market is CRIF, which stands for Centrale Rischi Finanziari and translates directly as "Centre (for) Financial Risks". The company was founded in Italy and became the first consumer bureau in 1988 in Bologna by a coalition of financial institutions, and it grew to encompass small and medium-sized firms. As of 2018, CRIF owned or controlled credit bureaus in 18 countries and provided solution assistance to 10 more privately operated bureaux and two public credit registries. Tiresias S.A. Banking Systems that serves the Greek market, has a cooperation with CRIF in multiple scorecard developments in the past.

2.4 Shortcomings Of Credit Scoring

Over the years, one of the biggest concerns on the field of credit scoring, is the accuracy of the models and their ability to predict or calculate efficiently and effectively the creditworthiness of the borrowers. This matter is probably the biggest concern when it comes to the loan contract between the lender and the borrower. For the lender is crucial that the credit score of the borrower is accurate, so that the lender does not undertakes higher risk than he can afford for his portfolio management and for the borrower is important as for the loan to be calculated fairly, which will result a fare interest rate of the loan or even get his credit application accepted.

Errors on credit reports were identified in two studies, one conducted in 2011 by an organization called PERC, which conducts research in collaboration with the CRAs.(Rosenblatt, 2020) The other was issued in 2013 by a regulator, the Federal Trade Commission. According to PERC (Policy and Economic Research Council), only 0.9 percent of credit reports contained inaccuracies that might lower credit scores by up to 25 points. The FTC discovered several minor issues, but just 2% of credit reports included mistakes that would result in a 25-point drop in credit score. The strategy for detecting inaccuracies was to recruit people who would evaluate their credit reports and object if they found an issue.

2.5 Credit Scoring Methods And Previous Research

Credit Scoring is a well-researched, due to the fact that credit risk assessment models find an important application in solving some of the biggest threats that institutions face, credit risk. Most of the recent studies in credit scoring are focus on the use of non “traditional” credit scoring methods. The latest research use Machine Learning models, such as “Decision Trees”(Wang et al., 2012) and “Non Linear Decision Trees”(Dumitrescu et al., 2022), “Random Forest”, “Extreme Gradient Boosting (XGBoost)” (Papilas, 2020). At the same path of Machine Learning methods, the “Lasso”, “Ridge”, “Elastic Net”, “Gaussian Naïve Bays”, “Support Vector Machine” (Wang et al., 2022), “Bagging”, “Adaboost”, “GBtree” and “Dart” techniques, were also have been tested.(Karezos, 2019). In another paper there have been developed a hybrid data mining model of feature selection and ensemble learning classification algorithms for credit scoring (Koutanaei et al., 2015). This research was proposing a three stage modeling that concluded that the classification results showed that the artificial neural network (ANN) adaptive boosting (AdaBoost) method has higher classification accuracy. Another paper proposed Credit scoring based on tree-enhanced gradient boosting decision trees (Liu et al., 2021; Liu et al., 2022)

Although plenty of studies have been concentrated to the models of credit risk assessment, not many studies have focused on the data for credit scoring and the sources of data ingestion. Since e-commerce is thriving and a lot of fintech companies offer solutions for “Buy Now Pay Later” products, a recent study focused on the online consumer lending and included a supplementary variables from alternative data sources and multilevel macroeconomic variables (Xia et al., 2021). Macroeconomic data about demographic segments as well as psychometric variables were also used (Djeundje et al., 2021) as for alternative data. Another study, was focused on sustainability credit risk assessment, mostly due to the environmental crisis that causes high financial risk to the

financial institutions and thus, this study incorporated sustainability data for the credit assessment (Zeidan et al., 2015). About the hotel sector there has been studied the inclusion of data from online booking platforms and social media, to credit risk modeling (Giannouli & Kountzakis, 2021)

2.6 Introduction To Machine Learning Methods For Credit Scoring

It is important to be clearly mentioned that the models that were developed and tested in this thesis are not focused on calculating the exact credit score of the "financial unit" that is being evaluated. The scope is to examine whether, the "alternative" data can work with Machine Learning methods for classification and inspect if the examined units have a good or bad credit behavior, or not. The methods that find an exact calculation of score, can later work as a binary classification method with some extra steps.

In this chapter, there is a concise introduction to the methods that were used to develop the alternative credit models of this thesis. The literature of these methods is vast and it could take a huge amount of pages to write about all these thoroughly. Though, explaining all the methods is not the scope of this thesis and thus there is just an introduction at this point.

2.6.1 Logistic regression

Nowadays, logistic regression is the most used approach for creating credit scoring models. Logistic Regression is used when the dependent variable is categorical. Through probability estimate, it aids in understanding the link between dependent variables and one or more independent variables. The logistic regression models are of a form of:

$$\mathbf{S = a + b_1x_1 + b_2x_2 +, \ldots, + b_kx_k}$$

The performance of models produced using logistic regression is often quite excellent with sufficient data pre-processing, and it is commonly regarded that logistic regression sets the bar against which other approaches for generating classification models should be judged. Given the default, logistic regression cannot be used to construct models for continuous objectives such as revenue, profit contribution, or loss. The maximum likelihood concept underpins logistic regression. The probability is computed as follows:

$$\mathbf{L= P_1 * P_2 *\ldots * P_G * (1 - P_{G+1}) * (1 - P_{G+2}) * \ldots * (1 - P_{G+B})}$$

- G represents the number of good items in the development sample.
- The number of bads in the development sample is denoted by B.
- P1,…, PG are the model's predicted probabilities of good for each good in the sample.
- The estimated probability of good for each bad in the sample are denoted by PG+1,…, PG+B.

- As a result, 1 – PG+1 represents the likelihood of bad for the first bad, 1 – PG+2 represents the probability of bad for the second bad, and so on.

The model's parameters are set to maximize the likelihood. Algorithms that iterate towards an optimum solution are used to find parameter coefficients. To begin, parameter coefficients are set to zero or randomly determined. The parameter coefficients are modified at each iteration of the algorithm depending on the change in probability observed from one iteration to the next. When the difference between iterations becomes negligible, the algorithm finishes. The Newton-Raphson technique is the most prevalent of these algorithms, and it is the default approach used by many major software products.(Finlay, 2010)

2.6.2 Decision Tree

Decision Trees are a sort of Supervised Machine Learning (you describe what the input is and what the related output is in the training data) in which the data is continually separated based on a certain parameter. Two entities may explain the tree: decision nodes and leaves. The decisions or consequences are represented by the leaves. And the data is separated at the decision nodes. There are two types of decision trees 1. Classification trees (Yes/No types) and 2. Regression trees (Continuous data types).(Myles et al., 2004). When there are missing features, a mix of category and numerical features, or a large variance in the size of features, they perform better in comparison to other methods.

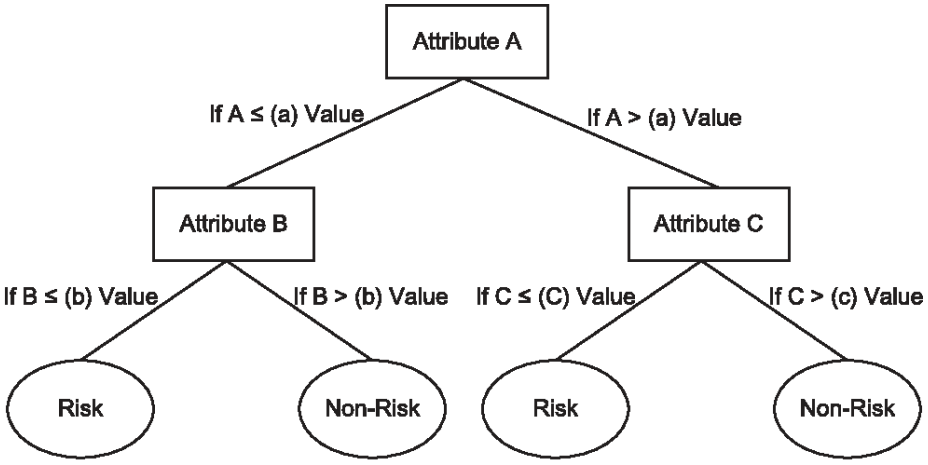


Image 1: Example of a decision tree in Credit Scoring application (Wang et al., 2012)

2.6.3 Gradient Boosting Tree

A Gradient Boosting Tree is a way for combining the results of many trees to do regression or classification. Both supervised and unsupervised learning use a large number of decision trees to limit the risk of overfitting (a statistical modeling error that occurs when a function is too tightly matched to a small number of data points, reducing the predictive potential of the model) that each tree faces alone. This approach uses Boosting, which involves sequentially adding weak learners

(usually decision trees with just one split, known as decision stumps) so that each new tree corrects the errors of the previous one.

The Gradient Boosting Algorithm is typically used to lower the Bias error, which is the amount by which a model's prediction differs from the target value. Gradient boosting is particularly effective when there are fewer dimensions in the data and there, a basic linear model performs badly, interpretability is not crucial, and there is no strict latency constraint.

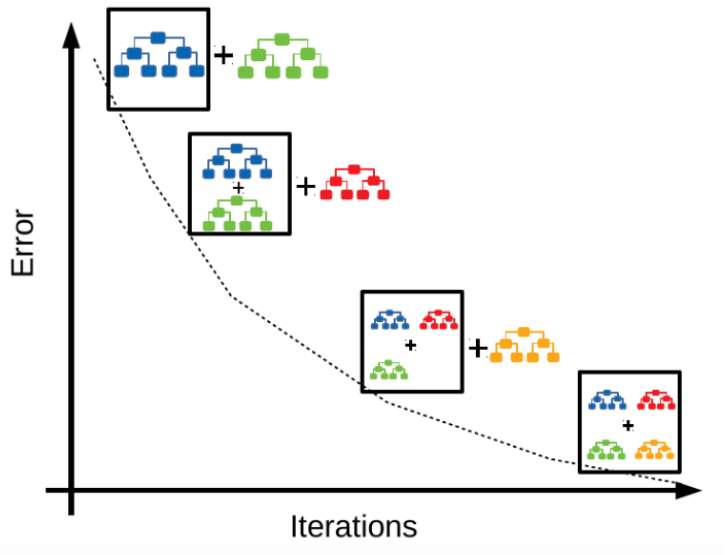


Image 2: Gradient Boosting Tree figure(Vasiloudis, 2019)

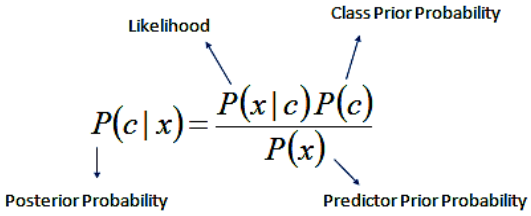
2.6.4 Random Forest

Random Forest is a method for resolving regression and classification problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. It consists of many decision trees, where the outcomes of every one of them will throw the final result taking the average or mean decisions. The greater the number of trees, the better precision of the outcome. Random Forest is appropriate when we have a huge dataset and interpretability is not a key problem, as it becomes increasingly difficult to grasp as the dataset grows larger. This algorithm is used in stock market analysis, diagnosis of patients in the medical field, to predict the creditworthiness of a loan applicant, and in fraud detection.

2.6.5 Naïve Bayes Classifiers

Naïve Bayes classifiers sit in the family of "probabilistic classifiers", which is the family of classifiers that are able to predict the probability of data, based on an input. Naïve Bayes classifiers assume that the data is independent of the value of all other data. It is made up of predictions based on the probability of items. It is dubbed Naive because it implies that the presence of one

feature is unconnected to the appearance of other features. This method is popular because it can outperform even the most advanced classification methods. Furthermore, it is simple to build and may be completed quickly. It is utilized to make real-time judgments because to its ease of use and efficiency. In addition, Gmail employs this algorithm to determine whether or not an email is spam.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Image 3: Bayes theorem

2.6.6 Support Vector Machine

Support-vector machines (SVMs, also known as support-vector networks) are supervised learning models with associated learning algorithms that examine data for classification and regression analysis in Machine Learning. AT&T Bell Laboratories created it. Given a series of training examples, each labeled as belonging to one of two categories, an SVM training algorithm constructs a model that assigns subsequent instances to one of the two categories, resulting in a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. (Suthaharan, 2016)

2.7 Accuracy Measures and Metrics

When it come to the evaluation of the models and the measure of data quality it is very important to discuss some of the most famous metrics that are being used in data science analysis. In this chapter, there are presented some of the most common metrics that were also used at the analysis phase of this thesis and based on them the results are commented.

2.7.1 Information Value

Although that the high computing power of the modern processors, allows the analysts to be a bit soft with the variable selection of the testing models, when the case is about big data with a lot of attributes, the analysis is not always very forgiving. In such cases, the computing time might be extremely long the processing workload heavy and the need to computing resources very expensive. To solve these problems at the stage of the analysis of the data set and the preprocessing

of the data, the analysts need to measure the predictive power of the independent variables and select the ones that they actually offer the best predictive results and exclude the ones that do not actually offer a lot to the models. One of the best metrics that offer this measurement is the Information Value or I.V. In other words Information Value provides a measure of how well a variable X is able to distinguish between a binary response (e.g. "good" versus "bad") in some target variable Y.(Osteyee & Good, 1974)

To see how this works, let X be grouped into n bins. Each x ∈ X corresponds to a y ∈ Y that may take one of two values, say 0 or 1. Then for bins Xi, $1 \leq I \leq n$,

$$\mathbf{IV}=\sum_{i=1}^{n}(gi - bi)* \ln(gi/bi)$$

Where,

b_i = (Number of 0's in X_i)/ (Number of 0's in X) = the proportion of 0's in bin i versus all bins

g_i = (Number of 1's in X_i)/ (Number of 1's in X) = the proportion of 1's in bin i versus all bins

The metric $\ln(g_i/b_i)$ is also known as the Weight of Evidence (for bin X_i).

2.7.2 Gini Impurity

Corrado Gini also devised a method for determining group homogeneity. Gini's impurity index is a straightforward calculation that is just the sum of the squared proportions (probabilities) that fall into each category. As a result, if there are two groups, there will be just two values (p and 1-p). For example, if there is a default rate of 5 percent the result would be 1-(0.052 + 0.952) = 0.095. The lower the result, the greater the homogeneity, such that zero means that all cases fall into a single category, and the maximum possible value means cases are uniformly distributed.

$$\textbf{Gini Impurity Index} \quad \mathbf{I = 1 - \sum_{j=1}^{g} \rho_j^2}$$

2.7.2 Confusion Matrices

A simple table comparing forecast vs actual is the Confusion Matrices. These tables show the answer on whether a prediction was correct or incorrect and are the most basic tool for assessing predictions. The confusion matrix states all possible combinations of test and truth for each of the classifications—True or False; Succeed or Fail etc. The confusion matrix lists all potential test and truth combinations for each classification

Table 2: A 2x2 confusion matrix

	Actual	
Predicted	Positive	Negative
Positive	TP	FP (Error Type II)
Negative	FN (Error Type I)	TN

Table 3: Performance Measures

Measure/ Also Called	**Calculation**
Sensitivity/ hit rate, recall	TP / (TP + FN)
Specificity	TN / (FP + TN)
False positive Rate / Fall-out Rate	FP / (FP + TN)
False negative Rate / Miss Rate	FN / (FN + TP)
Accuracy	(TP +TN) / (P+ N)
Pos. Predicted Value / Precision	TP / (TP + FP)
Neg. Predicted Value	TN / (TN + FN)
False discovery	FP / (FP + TP)

Positive or negative (yes or no) and true or false (right or incorrect) outcomes can be anticipated and real, with counts counted for each. Type I and Type II mistakes are incorrect Positive and Negative predictions. Which is worse varies, but while Positive is expensive, Type II's having Negative wrong failure to treat due to inaccurate diagnoses is also costly. Different ratios may be determined, the most essential of which are “sensitivity” and “specificity”, which are the ratios of correct Positive and Negative predictions.

2.7.3 Receiver Operating Characteristic (ROC)

The ROC has no identified author, most likely because it was the outcome of a highly covert war effort helped by MIT (Massachusetts Institute of Technology). After it was declassified, researchers at MIT and the University of Michigan worked on it further. It was used in engineering and psychology in the 1950s and 1960s to evaluate scarcely perceptible patterns. The ROC is widely utilized in medical, engineering, and other sectors today, including credit scoring.

The Receiver Operator Characteristic, or ROC curve was a visual representation of the proportion of true positives to false positives at different thresholds and basically is an evaluation metric for binary classification problems. Cases are classified in descending risk order in credit scoring, and the x- and y-axes are Goods and Bads, respectively. If one model's ROC curve is dominated (up and left) across the spectrum by another, it is the better model—lower errors at any cut-off. When the curves intersect, the dominant curve in the southwest corner is typically given preference.

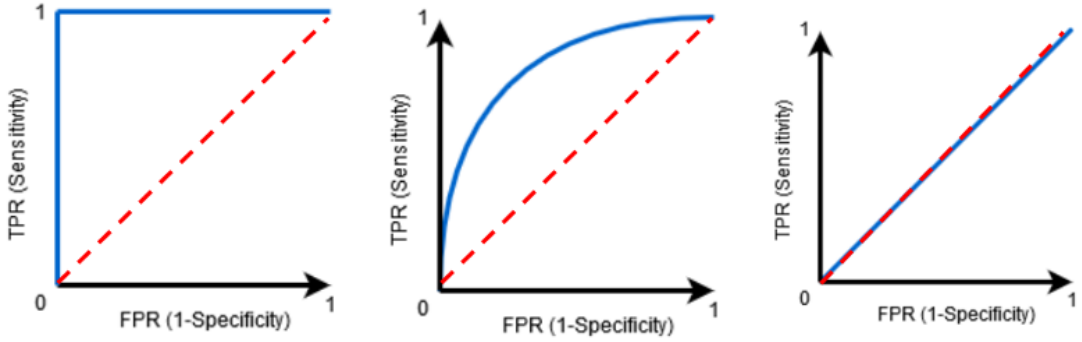


Image 4: Examples of ROC curves for different models

2.7.4 Area Under the Curve (AUC)

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. When AUC = 1, the classifier is capable of successfully distinguishing between all Positive and Negative class points. If the AUC was zero, the classifier would forecast all Negatives as Positives and all Positives as Negatives. (Bhandari, June 16, 2020)

When AUC1 is 0.5, there is a good possibility that the classifier will be able to discriminate between positive and negative class values. This is due to the fact that the classifier detects more True positives and True negatives than False negatives and False positives. When AUC=0.5, the classifier cannot differentiate between Positive and Negative class points. That is, the classifier predicts either a random class or a constant class for all data points. As a result, the greater a classifier's AUC score, the better its ability to discriminate between positive and negative classifications.

Gini vs AUROC $\mathbf{D_{AUC} = \frac{1+DGini}{2}}$ **, and** $\mathbf{D_{Gini} = 2 \times D_{AUC} - 1}$

2.7.5 SHAP Values

The Machine Learning models are sometimes referred to be "black-boxes," suggesting that you know the inputs and outputs but have little understanding of what is going on, behind all the calculations and the training of the models and the feature predictability. Recent papers have used SHAP Values in analyzing feature importance (Delgado et al., 2022) in Machine Learning models and the discriminative power of the features (Gramegna & Giudici, 2021).

SHAP Values or “SHapley Additive exPlanations” is a well-known subject in game theory and it was introduced by one of the 2012 Nobel Prize in Economic Sciences winners, Lloyd S. Shapley, who in 1953, came up with a solution for a cooperative game. Shapley wanted to calculate each player's contribution in a coalition game. Lets assume there are N players, and S is a subset of a number N of players. Let v(S) be the sum of the S players' values. Player i's marginal contribution upon joining the S players is $v(S \cup \{i\}) - v(S)$. If we take the average of the contributions across all potential coalition formation permutations, we get the contribution of player i.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

Image 5: SHAP Values equation

It is vital to note at this point that SHAP values do not show causality. The SHAP values can demonstrate how much each predictor contributes to the target variable, either positively or negatively. This is similar to the variable importance plot, except it may display the positive or negative association between each variable and the target.

3. Data Collection and Dataset Preprocessing

3.1 Alternative Data For Credit Scoring

Because the data landscape is ever-changing, analysts must update both their thinking and data collection methodologies to stay ahead of the curve. In many cases, data that was once thought unique, uncommon, or prohibitively expensive is now extensively used. Analysts that use these untapped data sources can obtain a competitive advantage before the rest of their sector does.

This type of data is known as alternative data, and with the ever-increasing amounts of data available in the modern world comes the possibility to get unique insights and competitive market edge. Alternative data can also be defined as data given from non-traditional sources; data that can be used to supplement traditional data sources to produce greater analytical insights that would not have been possible with traditional data alone. Simply said, it is data that is not generally used in a certain business but has the potential to be exploited to obtain a competitive advantage over others who do not have access to it.(Giannouli & Kountzakis, 2021)

3.2 Web-Scraping

Web scraping is a technique for extracting data from the World Wide Web (WWW) and saving it to a file system or database for subsequent retrieval or analysis. It is also known as web extraction or harvesting. Web data is commonly scraped using Hypertext Transfer Protocol (HTTP) or a web browser. This can be done manually by a person or automatically by a bot or web crawler. Because a massive quantity of heterogeneous data is continually created on the WWW, web scraping is generally recognized as an efficient and powerful tool for collecting big data.

Current online scraping techniques have evolved from smaller ad hoc, human-aided procedures to the use of fully automated systems capable of converting large websites into well-organized data sets in order to respond to a number of circumstances. Modern online scraping solutions can not only parse markup languages or JSON files, but they can also integrate with computer visual analytics and natural language processing to replicate how human users view web information.(Muehlethaler & Albert, 2021)

The process of gathering online resources and then extracting necessary information from the received data may be separated into two consecutive parts. A web scraping software, in particular, begins by creating an HTTP request to get resources from a certain website. This request can be structured as a URL with a GET query or as a chunk of HTTP message with a POST query. When

the request is successfully received and processed by the targeted website, the desired resource is obtained and returned to the web scraping software. The resource might be in a variety of forms, such as HTML web pages, XML or JSON data feeds, or multimedia data such as photos, audio, or video files. After downloading the online data, the extraction process proceeds to analyze, reformat, and organize the data in a systematic manner. A web scraping application must have two modules: one for creating an HTTP request, such as Urllib2 or selenium, and another for parsing and extracting information from raw HTML code, such as Beautiful Soup or Pyquery.

The Urllib2 module defines a set of functions for dealing with HTTP requests, such as authentication, redirections, cookies, and so on, whereas Selenium is a web browser wrapper that builds a web browser, such as Google Chrome or Internet Explorer, and allows users to programmatically automate the process of browsing a website. Beautiful Soup is intended for data extraction from HTML and other XML texts. It includes Pythonic methods for browsing, finding, and editing a parse tree, as well as a toolkit for dissecting an HTML page and extracting needed information using lxml or html5lib. Beautiful Soup can identify the encoding of the parser in progress and convert it to a client-readable encode automatically. Similarly, Pyquery has a collection of Jquery-like utilities. XML documents must be parsed. However, unlike Beautiful Soup, Pyquery only works with lxml for fast XML processing.

Web scraping may be used for a number of purposes, including contact scraping, price change monitoring/comparison, product review collection, real estate listing gathering, weather data monitoring, website change detection, and web data integration.

Although web scraping is an effective method for gathering big data sets, it is contentious and may generate legal issues about copyright, terms of service (ToS), and "trespass to chattels". A web scraper is allowed to copy data in figure or table form from a web page without violating copyright since it is difficult to prove copyright over such data because only a certain arrangement or selection of the data is legally protected. Regarding the Terms of Service, while most web apps incorporate some type of ToS agreement, its enforcement is typically ambiguous. For example, the owner of a web scraper that breaches the Terms of Service may claim that he or she never viewed or legally consented to the Terms of Service. Furthermore, if a web scraper sends too many data acquisition requests, this is functionally equivalent to a denial-of-service attack, in which the owner of the web scraper may be denied entry and liable for damages under the law of "trespass to chattels," because the owner of the web application has a property interest in the physical web server that hosts the application. By keeping a suitable querying frequency, an ethical web scraping application will avoid this issue. (Schintler & McNeely, 2019)

3.3The Collection Of The Alternative Data

Especially for the hotel sector that this thesis studies, there are plenty of studies that use data from well known commercial websites for reservation, such as “Booking.com” or “TripAdvisor.com”. Some thesis have also used data from social media for creating a rating model or measure the performance of a business. For this study it was chosen to develop a dataset with the alternative data was by a web scrapping program that was coded with the Python programming language. The

site that it was scraped is the website of the Hellenic Chamber of Hotels (HCH) with the weblink: https://www.grhotels.gr/en/.

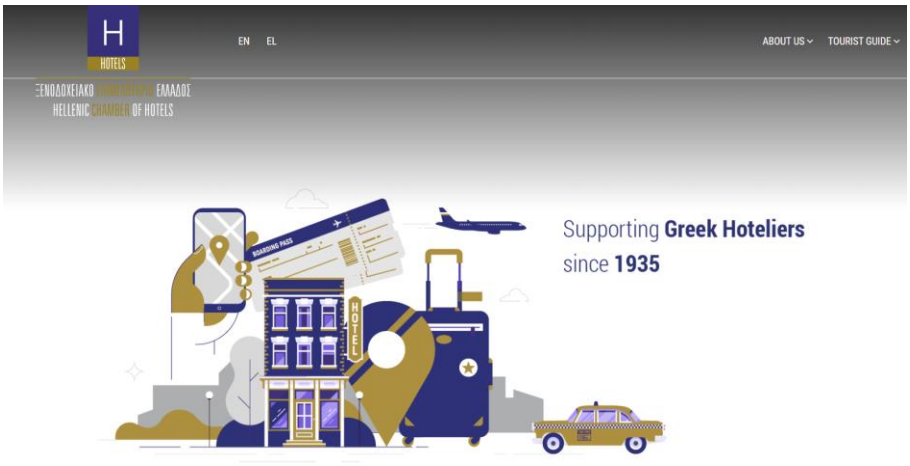


Image 6: The site of Hellenic Chamber of Hotels

The HCH operates since 1935 as a Legal Entity of Public Law. It is the institutional consultant of the Government as far as tourism and hospitality issues are concerned. Its members are, by law, all the hotels and camping sites of the country. It is run by an Administrative Council of elected representatives of hotels and camping sites as well as of representatives of the State. It is a member of the Confederation of National Associations of Hotels, Restaurants and Cafeterias of EU member-states (HOTREC). The Chamber's membership is about 10.000. Classical hotels are the most numerous.

It was not clear at the first that this site was the best source for alternative data. The stracture of the site was a major problem due to the difficulty to recognize the pattern for the scrapping application to run. Also it was not clear which values should been scrapped at first as there were a lot of missing values and the pages were not very well designed.

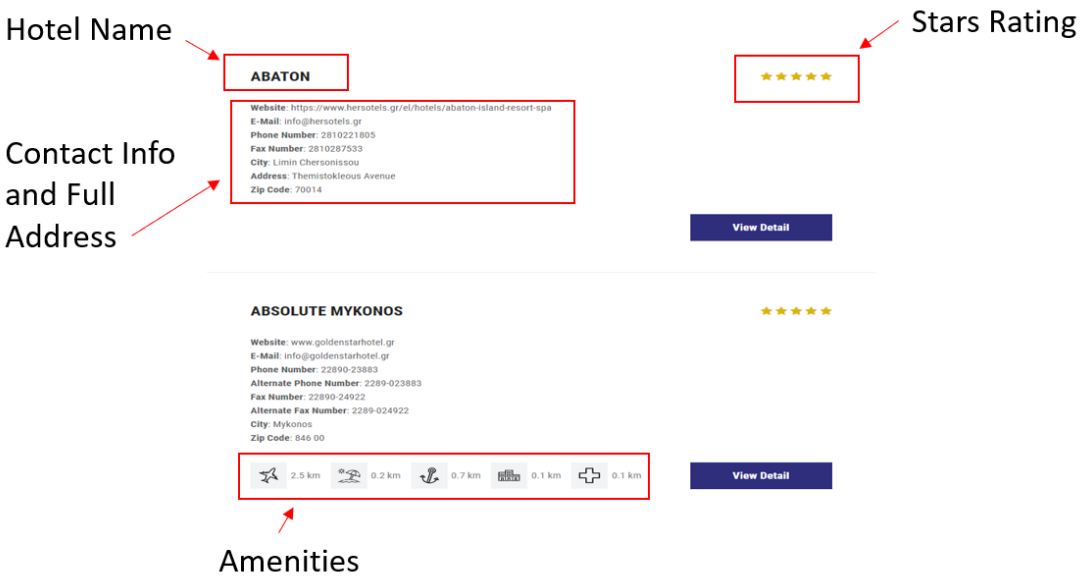


Image 7: A random search of hotels in the site

The code of the web scrapper was completely developed with Tiresias S.A. resources for research projects of the company and part of the code was also used at other projects of the company. That been said because of the company statement and permissions this code is not allowed for publication. But at this point it is worth mentioning that the code was developed with the Python programming language and there were also been used, “pandas”, ”re”, ”requests”, “BeautifulSoup” and “time” libraries. The coded needed 5 hours and 32 minutes to complete the scrape due to random time requests.

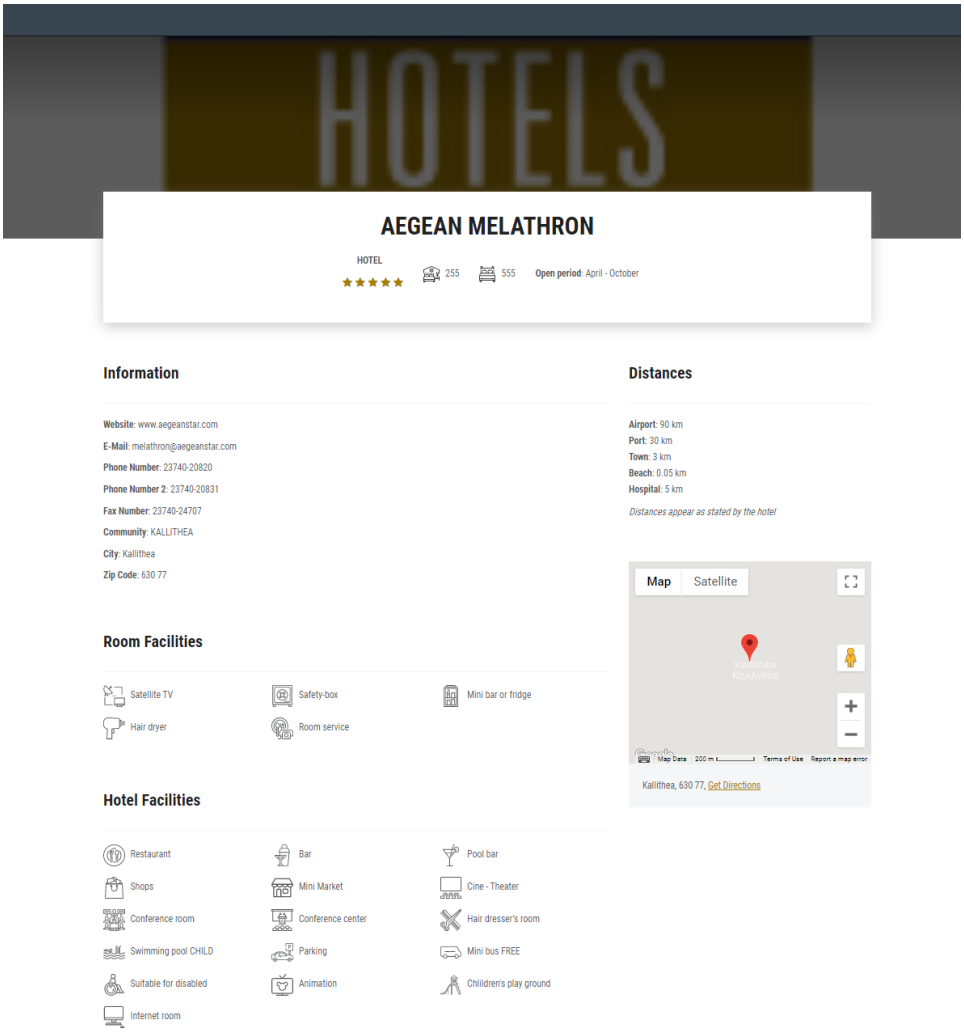
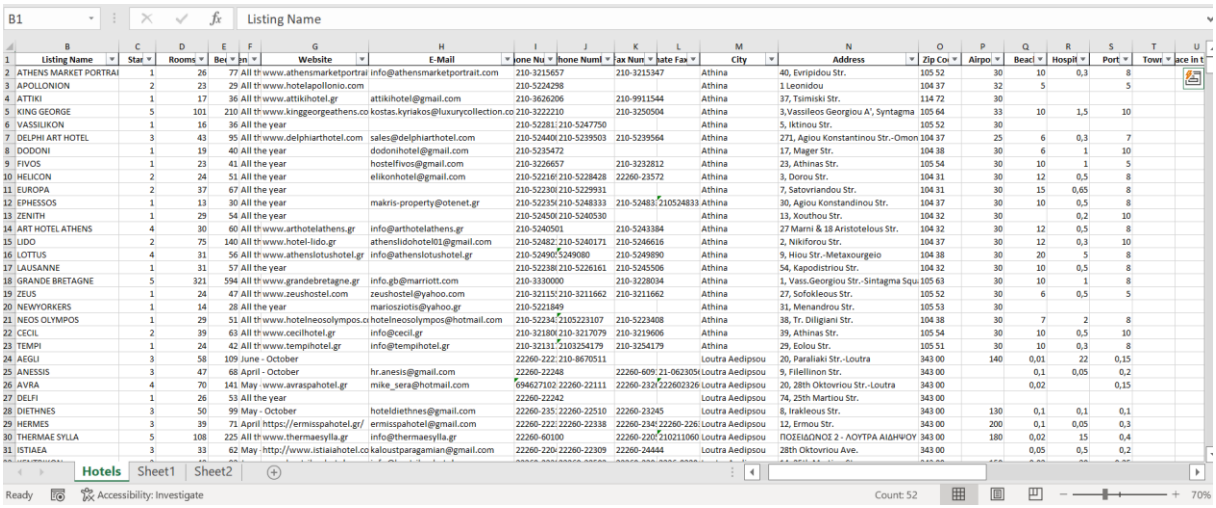


Image 8: The layout of the page that was scrapped, of a random hotel.

At the end of the execution of the code, the first dataset, that I named “Dataset-A”, of the data needed was created and it was saved as a .csv file. This dataset of hotels contains 10111 hotels, and data for 52 variables that consist of basic information for the hotel, and the room and hotel facilities. This is the dataset that gave the independent variables of the dataset, “X” values.



	Listing Name	Stars	Rooms	Beds	Open period	Website	E-Mail	Phone Number	Phone Number 2	Fax Number	Alternate Fax Number	City	Address	Zip Code	Airport	Beach	Hospital	Port	Town	Distance to the center
2	ATHENS MARKET PORTRAI	1	26	77	All th	www.athensmarketportrai	info@athensmarketportrait.com	210-3215657		210-3215347		Athina	40, Evripidou Str.	105 52	30	10	0,3	8		
3	APOLLONION	2	23	29	All th	www.hotelapollonio.com		210-5224298				Athina	1 Leonidou	104 37	32	5		5		
4	ATTIKI	1	17	36	All th	www.attikihotel.gr	attikihotel@gmail.com	210-3626206		210-9911544		Athina	37, Tsimiski Str.	114 72	30					
5	KING GEORGE	5	101	210	All th	www.kinggeorgeathens.co	kostas.kyriakos@luxurycollection.co	210-3222210		210-3250504		Athina	3,Vassileos Georgiou A', Syntagma	105 64	33	10	1,5	10		
6	VASSILIKON	1	16	36	All the year			210-52281	210-5247750			Athina	5, Iktinou Str.	105 52	30					
7	DELPHI ART HOTEL	3	43	95	All th	www.delphiarthotel.com	sales@delphiarthotel.com	210-52440	210-5239503	210-5239564		Athina	271, Agiou Konstantinou Str.-Omon	104 37	25	6	0,3	7		
8	DODONI	1	19	40	All the year		dodonihotel@gmail.com	210-5235472				Athina	17, Mager Str.	104 38	30	6	1	10		
9	FIVOS	1	23	41	All the year		hostelfivos@gmail.com	210-3226657		210-3232812		Athina	23, Athinas Str.	105 54	30	10	1	5		
10	HELICON	2	24	51	All the year		elikonhotel@gmail.com	210-52216	210-5228428	22260-23572		Athina	3, Dorou Str.	104 31	30	12	0,5	8		
11	EUROPA	2	37	67	All the year			210-52230	210-5229931			Athina	7, Satovriandou Str.	104 31	30	15	0,65	8		
12	EPHESSOS	1	13	30	All the year		makris-property@otenet.gr	210-52235	210-5248333	210-52483	210524833	Athina	30, Agiou Konstandinou Str.	104 37	30	10	0,5	8		
13	ZENITH	1	29	54	All the year			210-52450	210-5240530			Athina	13, Xouthou Str.	104 32	30		0,2	10		
14	ART HOTEL ATHENS	4	30	60	All th	www.arthotelathens.gr	info@arthotelathens.gr	210-5240501		210-5243384		Athina	27 Marni & 18 Aristotelous Str.	104 32	30	12	0,5	8		
15	LIDO	2	75	140	All th	www.hotel-lido.gr	athenslidohotel01@gmail.com	210-52482	210-5240171	210-5246616		Athina	2, Nikiforou Str.	104 37	30	12	0,3	10		
16	LOTTUS	4	31	56	All th	www.athenslotushotel.gr	info@athenslotushotel.gr	210-52490	5249080	210-5249890		Athina	9, Hiou Str.-Metaxourgeio	104 38	30	20	5	8		
17	LAUSANNE	1	31	57	All the year			210-52238	210-5226161	210-5245506		Athina	54, Kapodistriou Str.	104 32	30	10	0,5	8		
18	GRANDE BRETAGNE	5	321	594	All th	www.grandebretagne.gr	info.gb@marriott.com	210-3330000		210-3228034		Athina	1, Vass.Georgiou Str.-Sintagma Squ	105 63	30	10	1	8		
19	ZEUS	1	24	47	All th	www.zeushostel.com	zeushostel@yahoo.com	210-32115	210-3211662	210-3211662		Athina	27, Sofokleous Str.	105 52	30	6	0,5	5		
20	NEWYORKERS	1	14	28	All the year		mariosziotis@yahoo.gr	210-5221849				Athina	31, Menandrou Str.	105 53	30					
21	NEOS OLYMPOS	1	29	51	All th	www.hotelneosolympos.c	hotelneosolympos@hotmail.com	210-52234	2105223107	210-5223408		Athina	38, Tr. Diligiani Str.	104 38	30	7	2	8		
22	CECIL	2	39	63	All th	www.cecilhotel.gr	info@cecil.gr	210-32180	210-3217079	210-3219606		Athina	39, Athinas Str.	105 54	30	10	0,5	10		
23	TEMPI	1	24	42	All th	www.tempihotel.gr	info@tempihotel.gr	210-32131	2103254179	210-3254179		Athina	29, Eolou Str.	105 51	30	10	0,3	8		
24	AEGLI	3	58	109	June - October			22260-222	210-8670511			Loutra Aedipsou	20, Paraliaki Str.-Loutra	343 00	140	0,01	22	0,15		
25	ANESSIS	3	47	68	April - October		hr.anesis@gmail.com	22260-22248		22260-609	21-062305	Loutra Aedipsou	9, Filellinon Str.	343 00		0,1	0,05	0,2		
26	AVRA	4	70	141	May -	www.avraspahotel.gr	mike_sera@hotmail.com	694627102	22260-22111	22260-232	222602326	Loutra Aedipsou	20, 28th Oktovriou Str.-Loutra	343 00		0,02		0,15		
27	DELFI	1	26	53	All the year			22260-22242				Loutra Aedipsou	74, 25th Martiou Str.	343 00						
28	DIETHNES	3	50	99	May - October		hoteldiethnes@gmail.com	22260-235	22260-22510	22260-23245		Loutra Aedipsou	8, Irakleous Str.	343 00	130	0,1	0,1	0,1		
29	HERMES	3	39	71	April	https://ermisspahotel.gr/	ermisspahotel@gmail.com	22260-222	22260-22338	22260-234	22260-226	Loutra Aedipsou	12, Ermou Str.	343 00	200	0,1	0,05	0,3		
30	THERMAE SYLLA	5	108	225	All th	www.thermaesylla.gr	info@thermaesylla.gr	22260-60100		22260-220	210211060	Loutra Aedipsou	ΠΟΣΕΙΔΩΝΟΣ 2 - ΛΟΥΤΡΑ ΑΙΔΗΨΟΥ	343 00	180	0,02	15	0,4		
31	ISTIAEA	3	33	62	May -	http://www.istiaiahotel.co	kaloustparagamian@gmail.com	22260-220	22260-22309	22260-24444		Loutra Aedipsou	28th Oktovriou Ave.	343 00		0,05	0,5	0,2		

Image 9: Sample of the Dataset-A

Table 4: The list of the data variables that were scraped about general information for the hotels

Name of variable	Type	Description
Listing Name	string	Name of Hotel
Stars	integer	Number of Stars
Rooms	integer	Number of rooms
Beds	integer	Number of beds
Open period	string	Which period the hotel operates
Website	string	What is the website of the hotel
E-Mail	string	What is the email of the hotel
Phone Number	string	What is the phone number of the hotel
Phone Number 2	string	What is the alternative phone number of the hotel
Fax Number	string	What is the fax numberof the hotel
Alternate Fax Number	string	What is the alternative fax number of the hotel

City	string	The city where the hotel is located
Address	string	The adress where the hotel is located
Zip Code	integer	The postal code

Table 5: The list of the data variables that were scraped about distances from the closest locations

Airport	**decimal**	**How far is the hotel from an airport in KM**
Beach	decimal	How far is the hotel from an beach in KM
Hospital	decimal	How far is the hotel from a hospital in KM
Port	decimal	How far is the hotel from a port in KM
Town	decimal	How far is the hotel from a town in KM

Table 6: The list of the data variables that were scraped about the amenities

Fireplace in the rooms	**binary**	**If there exist (yes or no / binary 1 or 0)**
Hair dryer	binary	If there exist (yes or no / binary 1 or 0)
Mini bar or fridge	binary	If there exist (yes or no / binary 1 or 0)
Room service	binary	If there exist (yes or no / binary 1 or 0)
Safety-box	binary	If there exist (yes or no / binary 1 or 0)
Satellite TV	binary	If there exist (yes or no / binary 1 or 0)
Animation	binary	If there exist (yes or no / binary 1 or 0)
Bar	binary	If there exist (yes or no / binary 1 or 0)
Basketball	binary	If there exist (yes or no / binary 1 or 0)

Camping Card	binary	If there exist (yes or no / binary 1 or 0)
Casino	binary	If there exist (yes or no / binary 1 or 0)
Children's play ground	binary	If there exist (yes or no / binary 1 or 0)
Cine - Theater	binary	If there exist (yes or no / binary 1 or 0)
Conference center	binary	If there exist (yes or no / binary 1 or 0)
Conference room	binary	If there exist (yes or no / binary 1 or 0)
Garage	binary	If there exist (yes or no / binary 1 or 0)
Golf	binary	If there exist (yes or no / binary 1 or 0)
Hair dresser's room	binary	If there exist (yes or no / binary 1 or 0)
Internet room	binary	If there exist (yes or no / binary 1 or 0)
Mini bus FREE	binary	If there exist (yes or no / binary 1 or 0)
Mini Golf	binary	If there exist (yes or no / binary 1 or 0)
Mini Market	binary	If there exist (yes or no / binary 1 or 0)
Parking	binary	If there exist (yes or no / binary 1 or 0)
Pets allowed	binary	If there exist (yes or no / binary 1 or 0)
Pool bar	binary	If there exist (yes or no / binary 1 or 0)
Restaurant	binary	If there exist (yes or no / binary 1 or 0)
Roof garden	binary	If there exist (yes or no / binary 1 or 0)
Shops	binary	If there exist (yes or no / binary 1 or 0)
SPA-Thermal Baths	binary	If there exist (yes or no / binary 1 or 0)

Suitable for disabled	binary	If there exist (yes or no / binary 1 or 0)
Swimming pool CHILD	binary	If there exist (yes or no / binary 1 or 0)
Video - Pay TV	binary	If there exist (yes or no / binary 1 or 0)
Wifi	binary	If there exist (yes or no / binary 1 or 0)

In order for the models to be trained and tested, the credit scores of the hotels that already existed in the database of Tiresias S.A. because they have been scored with the traditional models and with strict financial data were retrieved. At this point, it is important to note that for this research a real-world credit scoring data set was taken from the private database of Tiresias S.A. and thus the data are protected and their handling follows the EU legal framework for Data Protection and all the directives of European Banking Authority and Bank of Greece.

From the database of the company, I collected and created a dataset with 3467 companies that run a hotel in Greece. The dataset, that I named "Dataset-B" was having data, about the name of the company, the year of last score, the address of the company along with the postal code and finally the score. The score variable of this dataset was the dependent variable of the models, the "Y" value.

At this point of the data handling, I had to face up a major problem. The two datasets did not have a primary key with which I could easily make the join of the two tables in order to connect the independent variables of the one set, with the corresponding values and their dependent variables of the second data set. At first it seemed that the best key to make the join, it would have been the VAT code, but Dataset-A did not have such an information and Dataset-B needed extra permissions to use these values. The second thought was to make the linkage based on the name o the hotel from Dataset-A and the name of the company from Dataset-B, but this brought very poor results. The vast majority of the companies from Dataset-B had the family name or just a very different name than the hotel's name that they actually connected to. For example, a hotel with the name "Blue Wave", was connected with the company "Papadopoulos Brothers" (after the transliteration process). These cases were more than 96% of the linkage tries.

The next thought was to use the addresses of the two dataset and make the join based on the matches there. That idea did not lead anywhere, as the values of both dataset where in different languages and the transliteration, the process of changing the alphabet characters between Latin characters and the Greek alphabet, created more problems than it solved. Also, a lot of addresses in Dataset-B were missing or mostly there were recorded incorrectly, or they were having different format than the one of the Dataset-A and that was the most often problem.

For example, there was a hotel with the address value as “25is Oktrobriou” after the transliteration at Dataset-B and the address at Dataset-A was recorded as “Eikostis Pemptis Oktobriou”. Another problem is that, a lot of streets can be found in different cities or even in multiple areas in the same city. That cause a lot of duplicate values. Last but not least, a small misspell of the address, will not bring very good results when simply statistical string comparing methods are used for data linkage. It is obvious that non o the record linkage programs can make a connection based on these data.

But when almost the last hope of a traditional record linkage manipulation of dataset was lost, an idea that later proved to work the best was born. The most famous search engines and their web maps application, show the same result to the user, when you try similar examples like the previous on, of alphabetical written streets and the same with arithmetic characters. The same results come out also when there are small misspells at the addresses. That is happening because the platforms connect the string of the address that is being searched, with a geographical coordinate instead of statistical methods for string analysis.

To test this hypothesis, I created a code that was reading the addresses of the datasets I wanted to join, then it was sending an API request to a platform that has expertise in geocoding and Geographical Information Systems databases, “Mapquest” https://developer.mapquest.com. Geocoding is the process, with which it can be found from a database which geographical coordinates connect to the geographical addresses that you search for. Mapquest offers some free API calls for geocoding application. With this code I got the longitude and latitude of the addresses of both companies and hotels from the two datasets. Basically, I transformed the strings of the geographical addresses, to geographical coordinates, with very good results.

Then I created new columns at the Dataset-A that mark an area of 50m by 50m and center the point of the geo coordinates that I have received from the GIS Database. Then I checked if a point from Dataset-B is inside the area that was mentioned before. This was done, because this GIS-dataset, has a slight deviation of the geo coordinates. If the point of the geo coordinates of Dataset-B was inside the marked of Dataset-A, then we could make a connection between the company and the hotel with a very good accuracy.

The next step was to remove the duplicate rows, the ones with missing values and the rows that corresponding to “special” categories based on their score. The duplicates and the missing values are creating a lot of problems on some of the Machine Learning models that were later tested and thus it was chosen to be excluded from the final dataset, in order for it to be tested with the same data for all the models. As for the special categories exclusions, this action was chosen as some scores indicate data anomalies that could influence the models negatively. The final dataset that was created after all these actions, contains of 223 rows, with all the values of the independent variables from Dataset-A and the scores as the dependent variables from Dataset-B. The final dataset will be called “Dataset-C”.

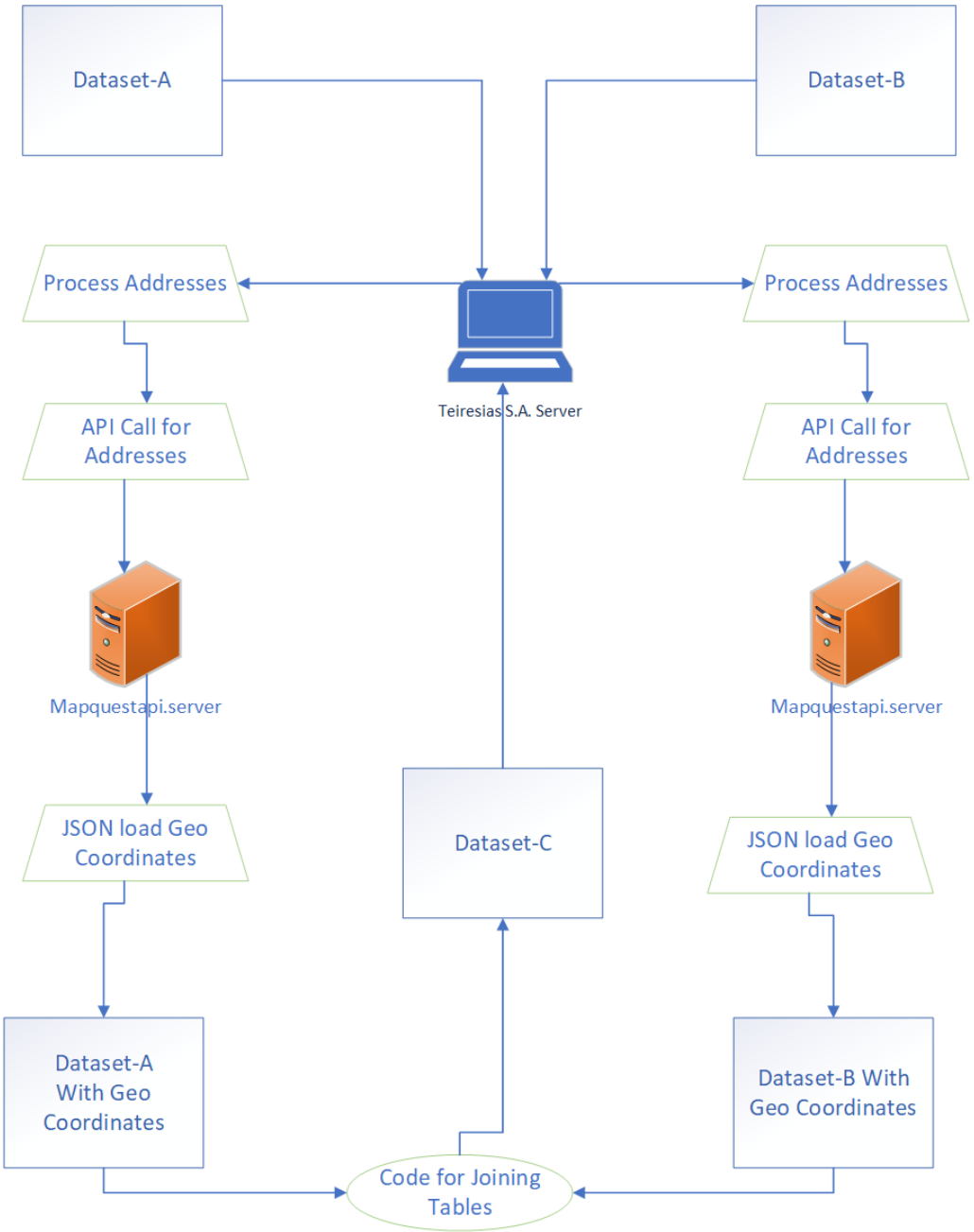


Image 10: Flow of data preprocessing

At this point it was chosen that there must be created some dummy variables in order to transform the continuous variables to categorical ones. This choice was made in order to facilitates the models to make simpler modeling by working with more discrete variables. A first intuitive choice was to set the distance between the hotels and the nearest town into 3 different categories. The first category (Town0) consisted of hotels that are closer to 1km away from a town, the second category (Town1) consist of hotels that are between 1km and 20km and the third category (Town2) consists of the hotels that are located in a distance more than 20km away from a town. There was not an equal way to distribute the hotels into categories that would not create more than 7 categories, thing that would probably cause other problems to the training of the models, such as

overfitting. So it was chosen to make categories that make sense as, “Hotel inside the City”, “Hotel very close to the city” and “hotel far from the city”.

For the other variables that were modified the metric of Information Values was used. By the use of this metric the continuous variables, were transformed to categorical. Also, the hotels with score that belong to the interval of [0, 350] or they have score “997” are characterized as “Bad” and the binary “1”. On the other hand, hotels with score between (350, 600], characterized as “Good” and the binary “0”.

Table 7: Score definition table for the variable of "Stars"

	Score_Definition						
Stars	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
1.0	14	8	22	36%	11%	8%	0,01
2.0	38	39	77	51%	30%	41%	0,04
3.0	40	28	68	41%	31%	29%	0,00
4.0	26	14	40	35%	20%	15%	0,02
5.0	10	6	16	38%	8%	6%	0,00
Total	**128**	**95**	**223**	43%	100%	100%	**0,07**

Table 8: Score definition table for the variable of “Room Bands”

	Score_Definition						
Room Bands	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
<=25	38	48	86	56%	30%	51%	0,11
>25	90	47	137	34%	70%	49%	0,07
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,18**

Table 9: Score definition table for the variable of “Bed Bands”

	Score_Definition						
Beds Bands	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
<=45	34	42	76	55%	27%	44%	0,09
46-194	70	44	114	39%	55%	46%	0,01
>=195	24	9	33	27%	19%	9%	0,06
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,17**

Table 10: Score definition table for the variable of "Open Period"

	Score_Definition						
Open period	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
All the year	56	53	109	49%	44%	56%	0,03
Not all year	72	42	114	37%	56%	44%	0,03
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,06**

Table 11: Score definition table for the variable of "Airport"

	Score_Definition						
Airport	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
missing	24	19	43	44%	19%	20%	0,00
<=80	93	60	153	39%	73%	63%	0,01
>80	11	16	27	59%	9%	17%	0,06
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,07**

Table 12: Score definition table for the variable of "Beach"

	Score_Definition						
Beach	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
missing	19	18	37	49%	15%	19%	0,01
<=0,30	65	38	103	37%	51%	40%	0,03
>0,30	44	39	83	47%	34%	41%	0,01
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,05**

Table 13: Score definition table for the variable of "Hospital"

	Score_Definition						
Hospital	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
missing	14	13	27	48%	11%	14%	0,01
<=7,00	91	74	165	45%	71%	78%	0,01
>7,00	23	8	31	26%	18%	8%	0,07
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,08**

Table 14: Score definition table for the variable of "Port"

	Score_Definition						
Port	**Good**	**Bad**	**Total**	**Bad Rate**	**good dist.**	**bad dist.**	**IV**
missing	65	39	104	38%	51%	41%	0,02
<=16	51	35	86	41%	40%	37%	0,00
>16	12	21	33	64%	9%	22%	0,11
Total	**128**	**95**	**223**	**43%**	**100%**	**100%**	**0,13**

The variables of "Room Bands" and "Bed Bands", along with the bands of "Port", are have the highest predictive power, out of the dummy variables that were created, as they present the higher IV metrics of 0.18, 0.17 and 0.13 respectively.

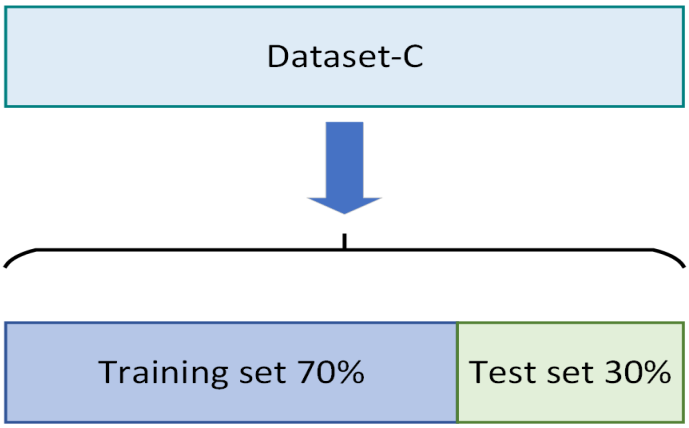


Image 11: The train and testing split of data

At the next step was to split the dataset into a training set and the testing set. The most common rate for the Train-Validation-Test through Cross Validation split is 70% for the training set and 30% for the test set. That split was conducted with randomly selected records from Dataset-C and in the end the training set consisted of 156 hotel records and the test set 67 records.

Table 15: Training and Test sets distribution of records and Good and Bad credit

Segment	**Number of Records**	**Good Credit**	**Bad Credit**
Train Set	156	86	70
Test Set	67	42	23
Total	223	128	95

Based on the previous tables, there were created the different categories that simplified the model development process. The next step was to develop the code of the model development phase and to examine how the models behave with the use of the alternative data and measure their accuracy

4. Results And Synopsis

4.1 Results

The final step of the analysis was to train the models and collect the result of the AUC metric of predictions on the test dataset. Based on the AUC we conclude the evaluation of models. The results are:

- Random Prediction: AUROC = 0.500
- Random Forest: AUROC = 0.618
- Naive Bayes: AUROC = 0.564
- Decision Tree: AUROC = 0.600
- Gradient Boosting: AUROC = 0.665
- K-Nearest Neighbour: AUROC = 0.515
- Logistic Regression: AUROC = 0.583
- Support Vector Machine: AUROC = 0.573

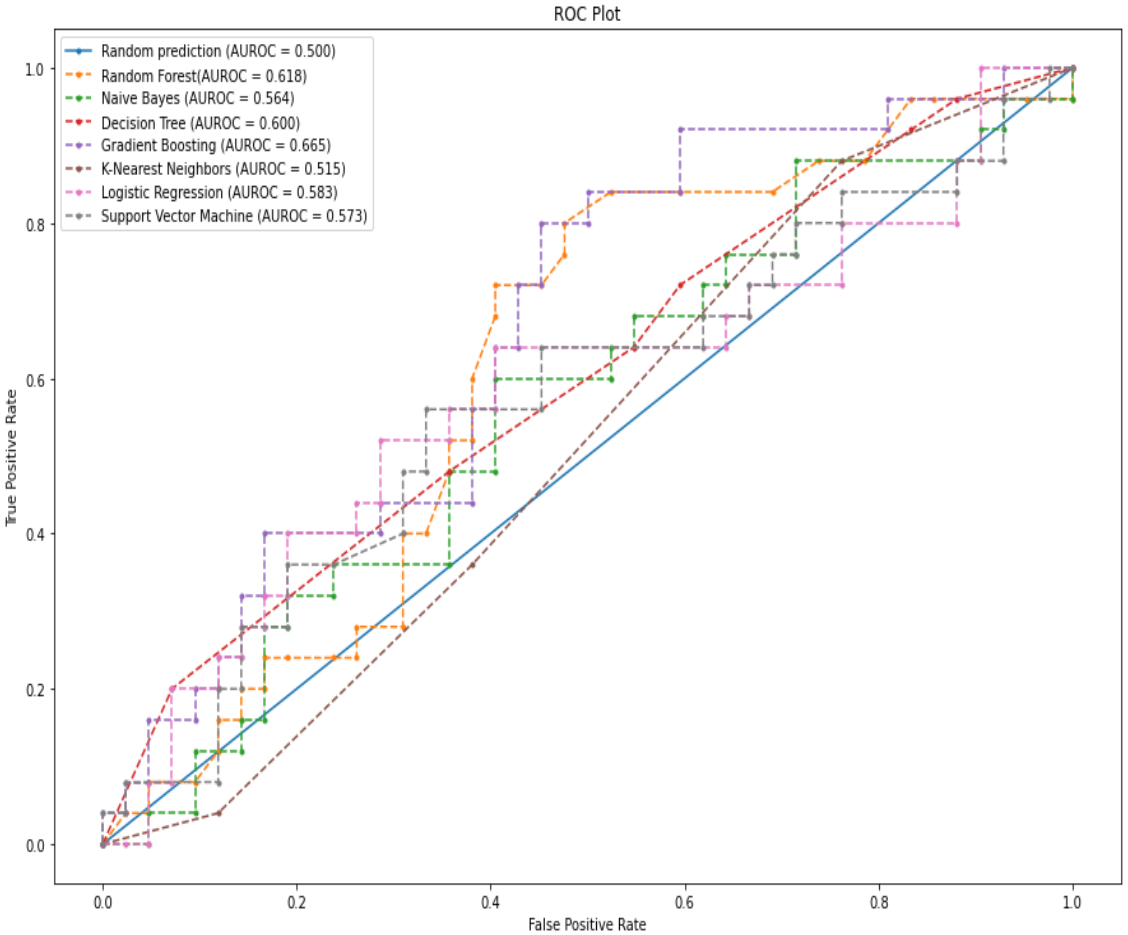


Image 12: The ROC curve of the models that were tested

As for the best model, the one that uses the Gradient Boosting classifier there where calculated the SHAP values of the features and the bar chart of contribution of the features is displayed on the following image.

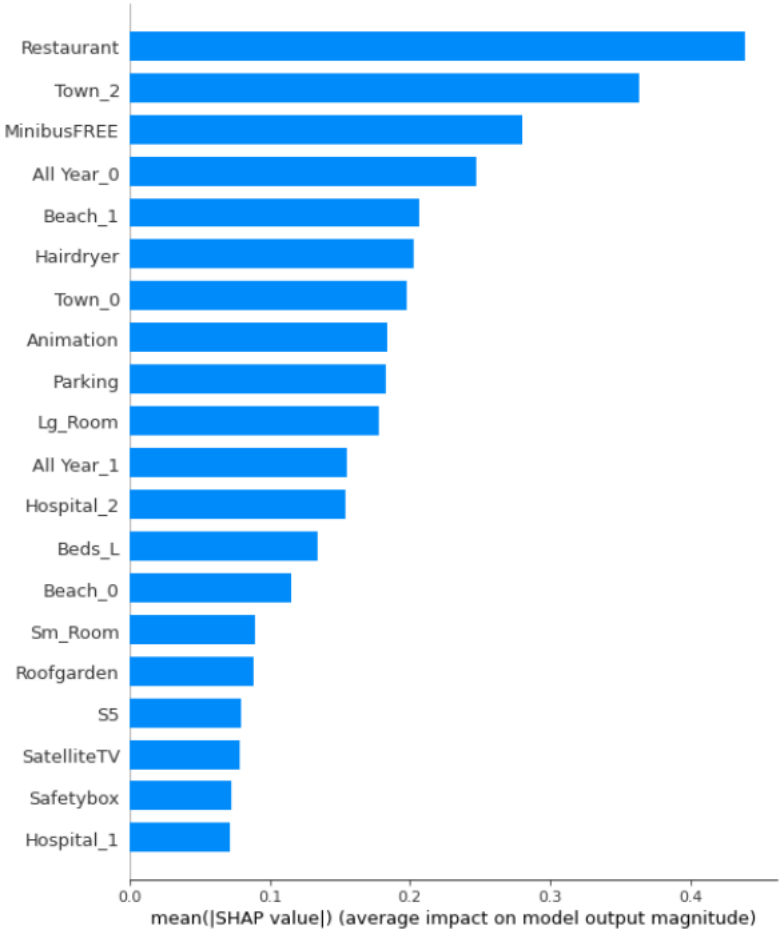


Image 13: Gradient Boosting Variable Importance Plot

4.2 Conclusion Comments Of the Analysis

Based on the AUC metric the model that best classifies the hotels to the ones with “Bad” credit scoring, so high probability to default and those ones who have “Good” credit scoring, so low probability to default, is the Gradient Boosting classifier with AUROC = 0.665. The second best model for the job with the AUROC = 0.618, uses the Random Forest classifier. That been said both of these models as long with the rest of the classifiers, do not present very accurate prediction results and they are having slightly more accurate predictions than the random predictor. This output was not a ground breaking event, due to the fact that in this research were used real life data that have not been tested before and so they have not proved any previous predictive power and test of them was actually one of the main purposes of the thesis.

This result may be correlated to various reasons that have to do with the tuning of the models but mostly with the data that were used as inputs. At the preprocessing stage there were removed a lot of records with null values, that could be manipulated with different data handling methods for

null values. This action may have caused irregularities in the dataset and possibly loss of information. The hypothesis for irregularities is based on the observation that a lot of hotels that link to records that were removed, are located in areas with low hotel density and in general smaller or less popular touristic destinations. Thus, records from Athens, Thessaloniki, Crete, Santorini, Rhodes and Mykonos, just to name a few, may have caused a qualitative imbalance at the dataset. Another problem has to do with the quantity of the data. In general Machine Learning models are having better training results when the input datasets are big. In this case, the sample of 223 records is a kind of a small percentage of the total population and a small dataset for the ML models to train and test the predictors.

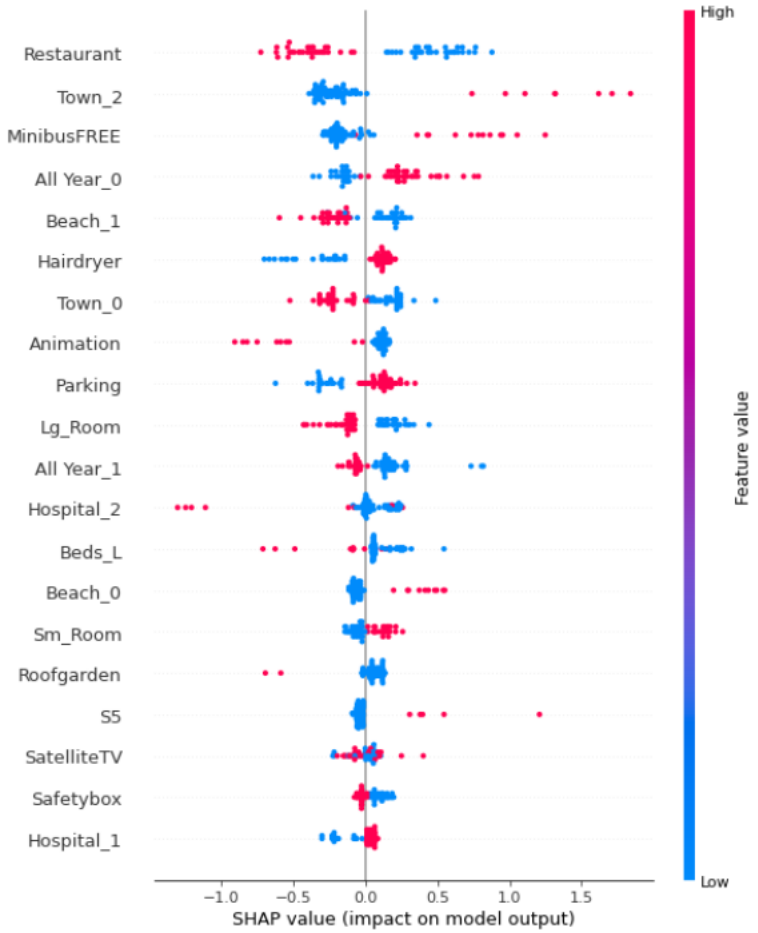


Image 14: Gradient Boosting Variables Summary Plot

From the SHAP values summary plot we can recognize the contribution of the predictors for the Gradient Boosting classifier, which was the model that gave the best predictive power based on the AUROC, the features that contributed the most at the model and also to see the positive and negative relationships of the predictors with the target variable. By this plot the variables are ranked in descending feature importance order, the horizontal location shows whether the effect of that value is associated with a higher or lower prediction, the color indicates whether the variable for that observation is high (in red) or low (in blue).

From the figure we can understand that the availability of restaurant, the location of the hotel and the offer of free mini bus by the hotel, are the variables with the biggest impact on the model. The existence of restaurant in the hotel, has a low positive impact for a bad credit classification. There is a low negative impact of the location of the hotels that are beyond 20 km away from a town, as well a low negative impact at the model of the feature of a free minibus amenity.

The variables that have the smallest role at the model, where the “Satellite Tv”, the “safety box” and the distance category between the hotel and the hospital, “Hospital_1”. An intuitive explanation is that, most hotels, good or bad credit scoring wise, have a Satellite Tv and safety box and they are also far from hospitals. So it is not strange that this variables have the lowest impact on the model, as they do not offer good distinction features to the model.

Although the analysis did not deliver results that could have been characterized as revolutionary or at least highly accurate, the journey of the analysis was extremely didactic, and it can be highlighted by two moments. The first has to do with the innovation of the analysis at the phase of the data processing and more specifically at the geocoding algorithm that later permit the join of the datasets. This process can be used in a wide variety of sets that cannot be joined otherwise. This simply idea of the geoinformation joining process will be used (by the writer of this thesis) in a lot of other projects that the typical record linkage algorithms are incapable to process. The second big moment is the one of the realizations of the great power of Machine Learning algorithms that can provide results, even without the desired accuracy, for better analysis in a relatively well-studied area, such that of credit risk analysis.

4.3 Future Work

Many aspects of research related to data science and its application to credit scoring were examined by this research. That been said, the end of this project and the results of it, may direct to some changes and new studies. Most of the different paths that could be chosen have to do with the data collection and processing, rather than the models.

Some of the ideas that could be tested differently, are:

1. Create segments of the dataset that are correlated with specific geographical areas and run individual models for each segment.
2. Include the segment with the missing values that was excluded from this analysis.
3. Create hybrid models that are based on different stages of modelling but still use alternative data.
4. Include alternative data from other sources, like social media presence, marketing budget of the hotels, and booking platform customer ratings.
5. The use of K-folding technique for choosing the training set and the test set for the model development and test.

Bibliography

Anderson, R. A. (2022). *Credit Intelligence & Modelling* (O. U. Press, Ed. 2nd ed.). https://doi.org/10.1093/oso/9780192844194.001.0001

Bhandari, A. (June 16, 2020). AUC-ROC Curve in Machine Learning Clearly Explained. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Delgado, P. C., Congregado, E., Golpe, A. A., & Vides, J. C. (2022). The Yield Curve as a Recession Leading Indicator. An Application for Gradient Boosting and Random Forest. *arXiv preprint arXiv:2203.06648*.

Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, *163*, 113766. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113766

Donaldson, T. H. (1989). *Credit Risk and Exposure in Securitization and Transactions*. Palgrave Macmillan. https://doi.org/DOI: 10.1007/978-1-349-10361-4

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192. https://doi.org/https://doi.org/10.1016/j.ejor.2021.06.053

Finlay, S. (2010). *Credit Scoring, Response Modelling and Insurance Rating*. Palgrave Macmillan. https://EconPapers.repec.org/RePEc:pal:palbok:978-0-230-29898-9

Giannouli, P., & Kountzakis, C. (2021). Data Analysis and Applications 4 Copyright Iste 2020 / File for personal use of Andreas Makrides only. In.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk [Original Research]. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.752558

Gravier, E. (2021). *Here's everything you need to know about your VantageScore credit score, plus how to check it*. https://www.cnbc.com/select/what-is-vantagescore/

Karezos, E. (2019). *Machine Learning Applications in Credit Scoring* Athens University of Economics and Business]. Athens.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*(C), 11-23. https://EconPapers.repec.org/RePEc:eee:joreco:v:27:y:2015:i:c:p:11-23

Liu, W., Fan, H., & Xia, M. (2021). Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Engineering Applications of Artificial Intelligence*, *97*, 104036. https://doi.org/10.1016/j.engappai.2020.104036

Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, 116034. https://doi.org/https://doi.org/10.1016/j.eswa.2021.116034

Muehlethaler, C., & Albert, R. (2021). Collecting data on textiles from the internet using web crawling and web scraping tools. *Forensic Science International*, *322*, 110753. https://doi.org/10.1016/j.forsciint.2021.110753

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, *18*(6), 275-285. https://doi.org/https://doi.org/10.1002/cem.873

Orgler, Y. E. (1970). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking*, *2*(4), 435. https://doi.org/10.2307/1991095

Osteyee, D. B., & Good, I. J. (1974). Information, weight of evidence, the singularity between probability measures and signal detection.

Papilas, K. (2020). *Credit Scoring through Machine Learning and Artificial Intelligence* ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS]. Athens.

Rosenblatt, E. (2020). *Credit Data and Scoring: The First Triumph of Big Data and Big Algorithms*. Academic Press.

Schintler, L. A., & McNeely, C. L. (2019). *Encyclopedia of big data*. Springer.

Suthaharan, S. (2016). Support Vector Machine. In (pp. 207-235). Springer US. https://doi.org/10.1007/978-1-4899-7641-3_9

Vasiloudis, T. (2019). Block-distributed Gradient Boosted Trees. http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowl. Based Syst.*, *26*, 61-68.

Wang, H., Chen, W., & Da, F. (2022). Zhima Credit Score in Default Prediction for Personal Loans. *Procedia Computer Science*, *199*, 1478-1482. https://doi.org/https://doi.org/10.1016/j.procs.2022.01.188

Xia, Y., Li, Y., He, L., Xu, Y., & Meng, Y. (2021). Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications*, *49*, 101095. https://doi.org/https://doi.org/10.1016/j.elerap.2021.101095

Zeidan, R., Boechat, C., & Fleury, A. (2015). Developing a Sustainability Credit Score System. *Journal of Business Ethics*, *127*(2), 283-296. https://doi.org/10.1007/s10551-013-2034-2

Βασικά Μεγέθη του Ελληνικού Τουρισμού 2019. SETE. https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/

Appendix

The code of the models developement

At this notebook I test some of the machine learning models for the classification problem for the Greek Hotel Sector

In [ ]:
```
import pandas as pd
import numpy as np
import seaborn as sb
np.random.seed(0)
import shap
```

In [ ]:
```
df1=  pd.read_excel(r'C:\Users\vgato\Desktop\code  diplomatikh\nocontval02.xlsx')
```

In [ ]:
```
df1.head()
```

Out[ ]:

0	Unnamed: 0	V1	Score_Definition	Hotel	Company	Stars	Fireplaceintherooms	Hairdrye
0	0	1022		CORALLI BLUE	2270	S2	0	
1	1	125		MANI	303	S2	0	
2	2	71		ITILO	154	S4	0	
3	3	265		MARATHEA	576	S3	0	
4	4	1308		XENONAS KAZAKOU	2917	S4	0	

5 rows × 46 columns

dhmiourgia dummy columns gia tis kolones poy exoyn kathgorikes metablhtes : Stars, Room_bands, Beds_bands, Open_Period_Indicator, Airport_Bands, Beach_bands Hospital_bands, Port_bands, Town_bands

```
StarsDummy                  =                  pd.get_dummies(df1['Stars'])
Room_bandsDummy=pd.get_dummies(df1['Room_bands'])
Beds_bandsDummy=pd.get_dummies(df1['Beds_bands'])
Open_Period_IndicatorDummy=    pd.get_dummies(df1['Open_Period_Indicator'])
Airport_BandsDummy=      pd.get_dummies(df1['Airport_Bands'])
```

```
Beach_bandsDummy=pd.get_dummies(df1['Beach_bands'])
Hospital_bandsDummy=pd.get_dummies(df1['Hospital_bands'])
Port_bandsDummy=pd.get_dummies(df1['Port_bands'])
Town_bandsDummy=pd.get_dummies(df1['Town_bands'])
```

In [ ]:
```
df1=   pd.concat((df1,StarsDummy,Room_bandsDummy,Beds_bandsDummy,Open_Period_Indicat
```

In [ ]:
```
df1
```

Out[ ]:

	Unnamed: 0	V1	Score_Definition	Hotel	Company	Stars	Fireplaceintherooms	Haird
1	1	125		MANI	303	S2	0	
3	3	265		MARATHEA	576	S3	0	
5	5	40		PANTHEON	85	S3	0	
7	7	1032		GLAROS	2295	S2	0	
8	8	456		TINION	1017	S3	0	
...	...	...	...	...	...	...	...	
408	408	990		LESVOS INN	2200	S3	0	
410	410	144		ATHENS MARKET PORTRAIT	347	S1	0	
411	411	416		KATALAGARI COUNTRY SUITES	910	S3	1	
413	413	832		MYTHOS	1842	S3	1	
414	414	42		KIPOS	89	S3	0	

223 rows × 73 columns

In [ ]:
```
df1 = df1.drop(['Stars','Room_bands','Beds_bands','Open_Period_Indicator','Airport
```

In [ ]:
```
df1.info()
```

Score_Definitions are edited / not displayed

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 223 entries, 1 to 414
Data columns (total 64 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            223 non-null    int64
 1   V1                    223 non-null    int64
 2   Score_Definition      223 non-null    object
 3   Hotel                 223 non-null    object
 4   Company               223 non-null    int64
 5   Fireplaceintherooms   223 non-null    int64
 6   Hairdryer             223 non-null    int64
 7   Minibarorfridge       223 non-null    int64
 8   Roomservice           223 non-null    int64
 9   Safetybox             223 non-null    int64
 10  SatelliteTV           223 non-null    int64
 11  Animation             223 non-null    int64
 12  Bar                   223 non-null    int64
 13  Basketball            223 non-null    int64
 14  CampingCard           223 non-null    int64
 15  Casino                223 non-null    int64
 16  Childrensplayground   223 non-null    int64
 17  CineTheater           223 non-null    int64
 18  Conferencecenter      223 non-null    int64
 19  Conferenceroom        223 non-null    int64
 20  Garage                223 non-null    int64
 21  Golf                  223 non-null    int64
 22  Hairdressersroom      223 non-null    int64
 23  Internetroom          223 non-null    int64
 24  MinibusFREE           223 non-null    int64
 25  MiniGolf              223 non-null    int64
 26  MiniMarket            223 non-null    int64
 27  Parking               223 non-null    int64
 28  Petsallowed           223 non-null    int64
 29  Poolbar               223 non-null    int64
 30  Restaurant            223 non-null    int64
 31  Roofgarden            223 non-null    int64
 32  Shops                 223 non-null    int64
 33  SPAThermalBaths       223 non-null    int64
 34  Suitablefordisabled   223 non-null    int64
 35  SwimmingpoolCHILD     223 non-null    int64
 36  VideoPayTV            223 non-null    int64
 37  S1                    223 non-null    uint8
 38  S2                    223 non-null    uint8
 39  S3                    223 non-null    uint8
 40  S4                    223 non-null    uint8
 41  S5                    223 non-null    uint8
 42  Lg_Room               223 non-null    uint8
 43  Sm_Room               223 non-null    uint8
 44  Beds_L                223 non-null    uint8
 45  Beds_M                223 non-null    uint8
 46  Beds_S                223 non-null    uint8
 47  All Year_0            223 non-null    uint8
 48  All Year_1            223 non-null    uint8
 49  Airport_0             223 non-null    uint8
 50  Airport_1             223 non-null    uint8
 51  Airport_2             223 non-null    uint8
 52  Beach_0               223 non-null    uint8
 53  Beach_1               223 non-null    uint8
 54  Beach_2               223 non-null    uint8
 55  Hospital_0            223 non-null    uint8
 56  Hospital_1            223 non-null    uint8
 57  Hospital_2            223 non-null    uint8
 58  Port_0                223 non-null    uint8
```

```
 59  Port_1                223 non-null    uint8
 60  Port_2                223 non-null    uint8
 61  Town_0                223 non-null    uint8
 62  Town_1                223 non-null    uint8
 63  Town_2                223 non-null    uint8
dtypes: int64(35), object(2), uint8(27) memory
usage: 70.3+ KB
```

In [ ]:
```
# x anexarthtes metablhites, y exartimeni metablhth

x = df1.values
```

In [ ]:
```
x
```

Out[ ]:
```
array([[1, 125, '', ..., 0, 0, 1],
       [3, 265, '', ..., 1, 0, 0],
       [5, 40, '', ..., 0, 1, 0],
       ...,
       [411, 416, '', ..., 0, 0, 1],
       [413, 832, '', ..., 0, 0, 1],
       [414, 42, '', ..., 1, 0, 0]], dtype=object)
```

In [ ]:
```
#delete merikes sthles poy den xreiazontan, mazi me thn sthlh exarthmenis metablht

x = np.delete(x,[0,1,2,3,4],axis=1)
```

In [ ]:
```
helper=df1

helper
```

Out[ ]:

	Unnamed: 0	V1	Score_Definition	Hotel	Company	Fireplaceintherooms	Hairdryer
1	1	125		MANI	303	0	1
3	3	265		MARATHEA	576	0	0
5	5	40		PANTHEON	85	0	1
7	7	1032		GLAROS	2295	0	0
8	8	456		TINION	1017	0	1
...	...	...	...	...	...	...	...
408	408	990		LESVOS INN	2200	0	1
410	410	144		ATHENS MARKET PORTRAIT	347	0	0
411	411	416		KATALAGARI COUNTRY SUITES	910	1	1
413	413	832		MYTHOS	1842	1	1
414	414	42		KIPOS	89	0	0

223 rows × 64 columns

In [ ]:
```
helper.drop(helper.columns[0], axis = 1, inplace = True)
helper.drop(helper.columns[0], axis = 1, inplace = True)
```

```
helper.drop(helper.columns[0], axis = 1, inplace = True)
helper.drop(helper.columns[0], axis = 1, inplace = True)
helper.drop(helper.columns[0], axis = 1, inplace = True)
```

In [ ]:
```
helper.shape
```

Out[ ]:
(223, 59)

In [ ]:
```
x.shape
```

Out[ ]:
(223, 59)

In [ ]:
```
from sklearn.model_selection import train_test_split

x_train ,x_test ,y_train ,y_test = train_test_split(x,y,test_size=0.3,random_state
```

Decision Tree Classifier

In [ ]:
```
#descision tree classifier

from sklearn import tree

dt_clf = tree.DecisionTreeClassifier(max_depth=5) #build
dt_clf.fit(x_train, y_train) #train
dt_clf.score(x_test,y_test) #make prediction
```

Out[ ]:
0.582089552238806

In [ ]:
```
from sklearn.metrics import confusion_matrix
(y_test, y_pred)
```

Out[ ]:
array([[27, 15],
[13, 12]], dtype=int64)

In [ ]:
```
y_pred = dt_clf.predict(x_train)
```

Out[ ]:
array([[81, 5],
[31, 39]], dtype=int64)

Random Forest Classifier

In [ ]:
```
from sklearn import ensemble

rf_clf = ensemble.RandomForestClassifier(n_estimators=100) #build

rf_clf.fit(x_train, y_train) #train
```

Out[ ]:
0.5970149253731343

Gradient Boosting Classifier

In [ ]:
```
gb_clf = ensemble.GradientBoostingClassifier()
gb_clf.fit(x_train, y_train) #train
gb_clf.score(x_test, y_test) #Make Prediction
```

Out[ ]:
```
0.582089552238806
```

Tune GB Classifier

In [ ]:
```
# Let's  tune this Gradient booster.

#gb_clf  =  ensemble.GradientBoostingClassifier(n_estimators=100)

#gb_clf.fit(x_train,y_train)
```

Naive Bayes Classifier

In [ ]:
```
from  sklearn.naive_bayes  import  GaussianNB
nb_clf       =       GaussianNB()       #build
nb_clf.fit(x_train,y_train)              #train
nb_clf.score(x_test,  y_test)  #Make  Prediction
```

Out[ ]:
```
0.6119402985074627
```

K-nearestneighboor

In [ ]:
```
from  sklearn.neighbors  import  KNeighborsClassifier
knn_clf     =     KNeighborsClassifier(n_neighbors=3)
knn_clf.fit(x_train,y_train)                  #train
knn_clf.score(x_test,  y_test)  #Make  Prediction
```

Out[ ]:
```
0.5223880597014925
```

Logistic Regression Classifier

In [ ]:
```
from  sklearn.linear_model  import  LogisticRegression
lr_clf             =            LogisticRegression()
lr_clf.fit(x_train,y_train)                    #train
lr_clf.score(x_test,  y_test)  #Make  Prediction
```

Out[ ]:
```
0.582089552238806
```

SVM Classifier

In [ ]:
```
from sklearn.svm import SVC

sv_clf=   SVC(probability=True,kernel='linear')
sv_clf.fit(x_train,y_train)              #train
sv_clf.score(x_test,  y_test)  #Make  Prediction
```

Out[ ]:
```
0.6268656716417911
```

Prediction Probabilities

In [ ]:
```
r_probs = [ 0 for _ in range(len(y_test))]
rf_prob    =    rf_clf.predict_proba(x_test)
nb_prob    =    nb_clf.predict_proba(x_test)
dt_prob    =    dt_clf.predict_proba(x_test)
gb_prob    =    gb_clf.predict_proba(x_test)
knn_prob   =   knn_clf.predict_proba(x_test)
lr_prob    =    lr_clf.predict_proba(x_test)
sv_prob  =  sv_clf.predict_proba(x_test)
```

probabilities positive outcome is kept

In [ ]:
```
rf_prob = rf_prob[:,1]
nb_prob = nb_prob[:,1]
dt_prob = dt_prob[:,1]
gb_prob = gb_prob[:,1]
knn_prob = knn_prob[:,1]
lr_prob = lr_prob[:,1]
sv_prob = sv_prob[:,1]
```

In [ ]:
```
knn_prob
```

Out[ ]:
```
array([0.33333333, 0.66666667, 0.        , 0.66666667, 0.33333333,
       0.        , 0.33333333, 0.33333333, 0.        , 0.33333333,
       0.        , 0.33333333, 0.66666667, 0.33333333, 0.33333333,
       0.        , 0.66666667, 0.66666667, 1.        , 0.33333333,
       0.        , 0.        , 0.66666667, 0.        , 0.        ,
       0.33333333, 0.33333333, 1.        , 0.66666667, 0.66666667,
       0.33333333, 0.        , 1.        , 0.66666667, 0.66666667,
       0.33333333, 0.33333333, 0.33333333, 0.33333333, 0.33333333,
       0.66666667, 0.33333333, 0.66666667, 0.33333333, 0.66666667,
       0.33333333, 0.33333333, 0.33333333, 0.33333333, 0.66666667,
       0.33333333, 1.        , 0.        , 0.        , 0.66666667,
       0.33333333, 0.66666667, 1.        , 0.        , 0.66666667,
       0.66666667, 1.        , 0.33333333, 0.33333333, 0.33333333,
       0.33333333, 0.66666667])
```

In [ ]:
```
from sklearn.metrics import roc_curve,roc_auc_score
r_auc = roc_auc_score(y_test,r_probs)
rf_auc =roc_auc_score(y_test,rf_prob)
nb_auc =roc_auc_score(y_test,nb_prob)
dt_auc =roc_auc_score(y_test,dt_prob)
gb_auc =roc_auc_score(y_test,gb_prob)
knn_auc =roc_auc_score(y_test,knn_prob)
lr_auc =roc_auc_score(y_test,lr_prob)
sv_auc =roc_auc_score(y_test,sv_prob)
```

Display the AUCROC Scores

In [ ]:
```
print("Random Prediction: AUROC = %.3f" %(r_auc))
print("Random Forest: AUROC = %.3f" %(rf_auc))
print("Naive Bayes: AUROC = %.3f" %(nb_auc))
print("Decision Tree: AUROC = %.3f" %(dt_auc))
print("Gradient Boosting: AUROC = %.3f" %(gb_auc))
print("K-Nearest Neighbor: AUROC = %.3f" %(knn_auc))
print("Logistic Regression: AUROC = %.3f" %(lr_auc))
print("Support Vector Machine: AUROC = %.3f" %(sv_auc))
```

```
Random Prediction: AUROC = 0.500
Random Forest: AUROC = 0.618
Naive Bayes: AUROC = 0.564
Decision Tree: AUROC = 0.600
Gradient Boosting: AUROC = 0.665
K-Nearest Neighbor: AUROC = 0.515
Logistic Regression: AUROC = 0.583
Support Vector Machine: AUROC = 0.573
```

Calculate the ROC Curve

In [ ]:
```
from sklearn import metrics
```

In [
```
# r_fpr, r_tpr, _ = roc_curve(y_test,  r_probs,pos_label=0)
# rf_fpr, rf_tpr, _ = roc_curve(y_test,  rf_prob,pos_label=0)
# nb_fpr, nb_tpr, _ = roc_curve(y_test,  nb_prob,pos_label=0)
# dt_fpr, dt_tpr, _ =roc_curve(y_test,  dt_prob,pos_label=0)
# gb_fpr, gb_tpr, _ = roc_curve(y_test,  gb_prob,pos_label=0)
# knn_fpr, knn_tpr, _= roc_curve(y_test,  knn_prob,pos_label=0)
```

In [ ]:
```
y_test.astype(int)
```

Out[ ]:
```
cannot display score definition--- Nikolas Gatos Edit
```

In [ ]:
```
r_fpr
```

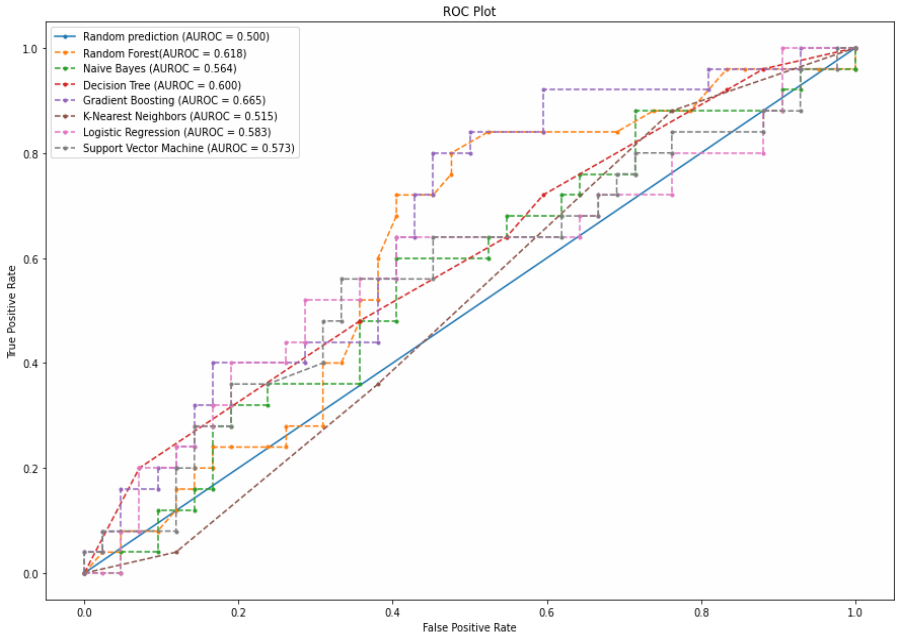
Out[ ]:
```
array([0., 1.])
```

In [ ]:
```
import matplotlib.pyplot as plt
# plot the roc curve for the model


#gia megalytero plot
plt.figure(figsize=(15,10))


plt.plot(r_fpr, r_tpr, marker='.', label='Random prediction (AUROC = %0.3f)' %r_au
plt.plot(rf_fpr, rf_tpr, linestyle='--', marker='.',label='Random Forest(AUROC = %
plt.plot(nb_fpr, nb_tpr, linestyle='--', marker='.',label='Naive Bayes (AUROC = %0
plt.plot(dt_fpr, dt_tpr, linestyle='--', marker='.', label='Decision Tree (AUROC =
plt.plot(gb_fpr, gb_tpr, linestyle='--', marker='.', label='Gradient Boosting (AUR
plt.plot(knn_fpr, knn_tpr, linestyle='--', marker='.', label='K-Nearest Neighbors
plt.plot(lr_fpr, lr_tpr, linestyle='--', marker='.',label='Logistic Regression (AU
plt.plot(sv_fpr, sv_tpr, linestyle='--', marker='.', label='Support Vector Machine


#Title
plt.title('ROC Plot')
```



Plot the ROC Curve

In [ ]:

```
dt for
```

Out[ ]:

```
array([0.        , 0.07142857, 0.35714286, 0.54761905, 0.5952381 ,
       0.83333333, 0.88095238, 1.        ])
```

Shap Values for Gradient Boosting

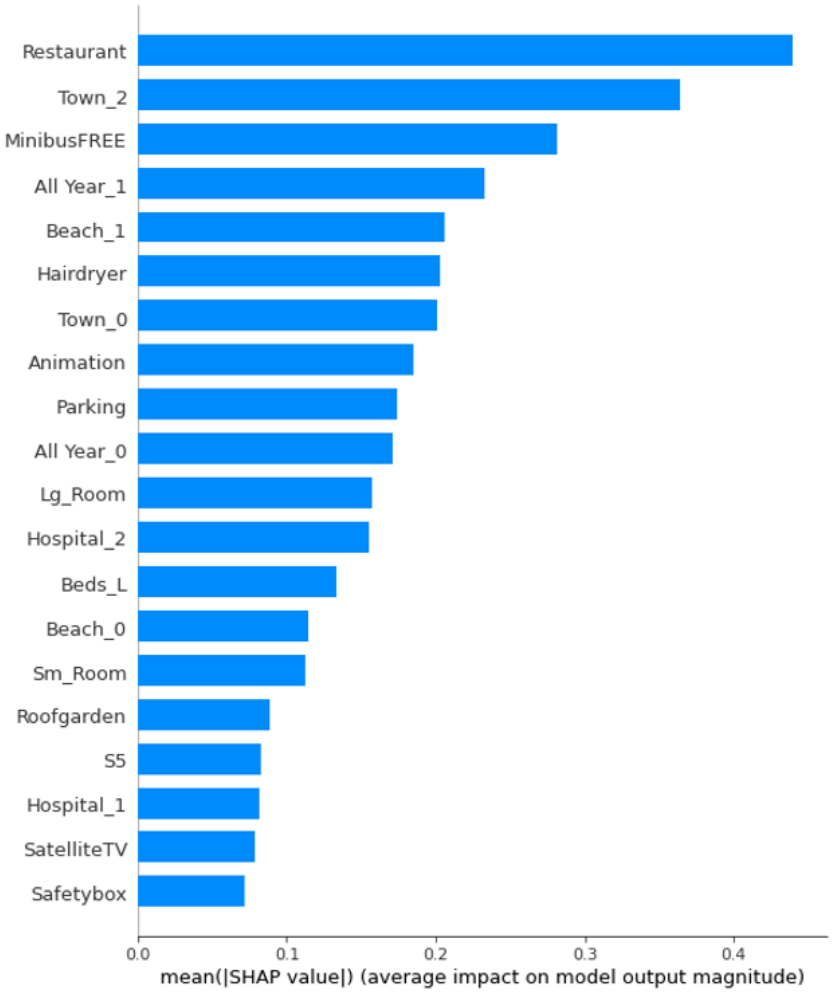
In [ ]:

```
gb_explainer = shap.Explainer(gb_clf)
gb_shap_values = gb_explainer(x_test)
```

In [ ]:

```
#shap.summary_plot(gb_shap_values, x_test, plot_type="bar")

shap.summary_plot(gb_shap_values, helper.columns, plot_type="bar")
```



In [ ]:
```
shap.summary_plot(gb_shap_values, x_test, helper.columns)
```