****

**Security of E-based System**

**Project**

**CREDIT RISK ANALYSIS**

**Submitted by:**

**Moukhik Misra - 19BCE2190**

**Abstract**

Credit scoring is probably one of the most "modern" predictive modelling techniques, to determine whether or not credit provided to an individual would possibly result in the lending institution's benefit or loss. There are several differences and nuances as to how precisely credit is provided to individuals , companies, and other organisations for different purposes (purchasing equipment, real estate, consumer goods, etc.) and using different credit systems (credit card, loan, deferred payment scheme). In all situations, however, a lender provides an individual or entity with money and expects to be paid back in time with interest in proportion to the risk of default.

A lender makes two types of decisions commonly: first, whether to give a new applicant credit, and second, how to deal with current applicants, including whether to lift their credit limits. In both cases, whatever the methods used, it is important to have a wide sample of prior customers available with descriptions of their application, behavioural trends and subsequent credit history. Most techniques use this sample to identify the relation between consumers' characteristics (yearly income, age, number of years of employment with their current employer, etc.) and their subsequent history. We attempt to implement a similar technique to achieve more efficient analysis of credit risk management.

Some of the methods proposed by us are: 1. Data preprocessing by missing values treatment (Numerical and Catergorical) and Outlier and Frequency Analysis. 2. Data analysis by use of T-Test and Data Visualization by use of graphs and heat maps. 3. Variable Transformation done by P-value from above test. 4. Variable Reduction done by Multicollinearity analysis (by VIF factor). 5. Implementing Logistic Regression by using Statistical Model 6.Getting the confusion matrix. 7. Decile analysis to be done to verify the statistical model 8.Getting the accuracy of the model.

We use Jupyter notebooks to implement the code and get a more precise and step-by-step analysis of the result. Jupyter notebook is part of the anaconda suite of python tools which enable us to visualize data to a great degree with quite a lot of ease. It also provides the tools for machine learning and is quite easy to debug as the entire code is compiled line by line and can be divided into subsections as per the programmer’s choice.

**Introduction**

A credit risk is the danger of default on a loan that could result from a borrower failing to meet the necessary payments. In the first resort, the danger is that of the lender which entails loss of principal and interest, interruption of cash flows and increased cost of collection. The loss may be absolute or minor. Higher credit risk levels would equate with higher borrowing rates in an efficient economy. Because of this, borrowing cost metrics such as yield spreads can be used to infer credit risk levels based on market participants' assessments.

To minimise the credit risk of the lender, the lender may carry out a credit check on the prospective borrower, can enable the borrower to take out sufficient insurance, such as mortgage insurance, or may request protection on any of the borrower's properties or a third party guarantee. The lender may also take out risk insurance, or sell the debt to another company. In general, the higher the risk, the higher the rate of interest the debtor is required to pay on the debt. Credit risk occurs primarily when investors knowingly or unwillingly fail to pay due.

Most lenders use their models (credit scorecards) to rank potential and existing customers by risk, and then apply suitable strategies. Lenders charge a higher price for higher-risk customers with products such as unsecured personal loans or mortgages, and vice versa. By setting credit limits the risk is controlled with revolving products such as credit cards and overdrafts. Some products also require collateral, normally an asset pledged to secure the loan's repayment.

Credit rating models also form part of the process that banks or lending agencies use to offer customers credit. These models typically have qualitative and quantitative sections for corporate and commercial borrowers detailing different aspects of the risk including, but not limited to, operational experience, management skills, asset quality, and leverage and liquidity ratios, respectively. Once this information has been thoroughly checked by credit officers and credit committees, the lender will supply the funds in compliance with the terms and conditions of the contract.

Several methods have been proposed to develop a credit risk analysis model which can accurately predict the likelihood of a person defaulting on loans. Almost all of these methods rely on machine learning techniques using which a model is created and trained by giving it data from datasets and it is then tested to check for its accuracy.

The main objective is to obtain the greatest possible accuracy as this results in the greatest profits for the lenders, which are mostly banks. An accurate model could be the difference between profits and millions of dollars of loss for an organization. All the new age credit risk scoring models rely on machine learning techniques such as Logistic Regression, Naïve Bayes, Knn, Random Forest and such models.

Nowadays there are several machine learning classifiers which can be used to provide an effective and accurate prediction model. Among these model, Logistic Regression is an extremely popular and accurate model which is widely used for modelling applications. Logistic Regression is based on the principle of Log Odds. Odds is the probability if an event occurring.

Logistic Regression is a bi-variate classifier which means that it compares the relationship between two variables and determines the impact that the change in one variable causes a change in the other. Logistic Regression is a light classifier model ad hence it is quite space and time efficient which is an important factor in deciding which model is going to be used in the corporate world.

Before applying classifiers however, it is essential to qualitatively view the data that is bein given to the model for training purposes. Hence data visualization and data exploration form two important factors when working with large amounts of data. This is used to see the comparisons between the different parameters of data in a qualitative sense. Data visualization for a large amount of data must be simple and clean so that people are easily able to understand the trends and take the necessary steps manually.

Histograms are extremely simple and basic to understand and they are very popular form of data visualization. Boxplots are also easy to understand graphs which can be used to compare the mean values of different parameters and hence look at the trends and comparisons between the parameters.Multiple classifier tests are used to determine the relations between the variables. Statistical tests are used to determine relations between parameters. Although visualizing the data gives us a qualitative understanding and a basic over view of the trends, the real quantitative data for relations between variables can be obtained only by using Statistical models such as t-test and VIF.

History itself provides the best motivations to enhance the research in the credit risk analysis and get accuracy in figures behind the lender-borrower relationship. The Global Economy crisis of 2008 saw a bubble in the real estate sector whose causes can be traced back to unmonitored loan providing in the major banks in United States of America. There is no doubt that bad analysis of Credit Risk in banks can lead to unimaginable lossses in bussinesses and destruction of even the major economies in the world. The large scale destruction potential of these activities are often underestimated, though these issues should and must concern everyone and not just economists since it holds the potential to push most of the countries in deep crisis like we've seen before.

Therefore, we in this project have given a miniature effort to bring more efficiency to the calculation of credit risk using visual and graphical representations to present a more precise and yet understable form of data representation. Using stastical models and concepts to present the information have always been very convinient. It definately has been represented even better after we took each variable's seperate visualisation for each form of graph. Thus, we attempted to use them to graphically present the concepts of Credit Risk Analysis and explain the figures and conclusions better.

**Motivation**

With so many of the world's largest lending institutions collapsing after the financial crisis of 2007, companies have begun realising the vital value of credit risk management. Without careful supervision, they realised exposure to risk could prove devastating.

To start managing credit risk you need to know what it is first. Credit risk refers to the possibility of a counterparty not repaying an organisation 's debt. This means risk may be true, non repayment realised, or potential non-performance. The corporation owing money loses income when a counterparty defaults on a loan.

Enterprise-level credit risk analysis helps executive management and risk practitioners to consider the potential accounts may come at too high a risk and beyond their defined risk tolerance. Credit ratings include an individual scale to measure the organisations' credit worthiness and riskiness. The credit rating firms named "Big Three" are Moody's, S&P and Fitch. The ratings range from AAA, prime and riskless investment, to DDD, a default trash. Any ranking above BBB- shall be deemed to be investment grade.

Organizations that prioritise risk management aim to assess the amount of risk they would be able to handle in a worst case scenario. The OATI Commodity Trading and Risk Management (CTRM) solution, OATI web CTRM, effectively decides acceptable levels of risk, with a front-to-back office suite that monitors and identifies credit risk that could endanger your company, allowing you to concentrate on the revenues.

The OATI solution measures current and potential exposures, assesses and monitors criteria for inbound / outbound collateral, produces internal credit ratings, and more. Our automated system basically does the heavy lifting in credit risk management, so you don't have to.

**Aim**

Credit scoring is the compilation of decision models and their underlying techniques that assist consumer credit lenders in granting. These techniques decide who will get credit, how much credit they need to get, and what organisational methods will increase the lenders' profitability. In addition, they aid in the evaluation of lending risk. Credit scoring is an accurate measure of a person's credit rating, as it is based on real data. Typical consumer market applications include: credit cards, auto loans, home mortgages, home equity loans, mail-catalog orders, and a broad range of personal lending products. We collect details about various credit of various customers and attempt to develop a python program to predict the validity of credit risk of the customer.

**Objective**

Credit management is not about finding the best way, the most efficient way possible, to minimise debt. It's about building trustworthy relationships with customers to achieve business outcomes and to increase profits.

Safeguarding consumer risk, settling unpaid balance sheets and maximising cash flow are three main credit management priorities that are crucial to find sustainable success.

Using empathy and honesty, these three main values are employed to produce optimal performance. We are all human beings, after all, on a quest to do the best we can. Cash management will be of benefit to us.

And most importantly, have fun while doing this project.

**Literature Review**

**[1] Credit Risk Evaluation Model with Textual Features from Loan Descriptions for P2P Lending**

The main aim of credit risk evaluation is to find the probability that a loan borrower will default or be overdue on a loan. Before the advent of modern technologies, traditional statistical methods such as discriminant analysis was used for credit risk evaluation. However, with the development of powerful Machine Learning techniques, techniques such as SVM, neural network and regression methods in addition to deep learning have taken over the scene. Nowadays almost all credit evaluation models use machine learning techniques such as Logistic Regression, artificial neural network, naïve bayes, SVM decision tree etc. With the development of big data analysis techniques, some complex classifiers such as deep learning techniques have been utilized to build accurate risk evaluation models.

Deep learning is considered an important solution to the manual labelling of conventional machine learning approaches in the age of big data and is also an important tool in big data analytics. However, a large number of recent studies have concentrated on classical machine learning approaches in the field of credit risk assessment, and few have used deep learning to determine credit risk, the latter being attributable to its low explanatory capacity and high training difficulty. The future usefulness of deep learning methods in credit risk assessment in respect to P2P lending is promising.

Multiple text feature extractors, namely, the LDA topic model, CNN, EA, and the TE, have been used to predict loan defaults [1]. The Latent Dirichlet Allocation-topic model is intuitively a simple model of text classification, which aims to categorise each text into one or more of the predefined categories; in this model, each category can be viewed as a latent subject. The Convolutional Neural Network extracts all the important local characteristics of each text, and the Embedding Average simply returns the core of each text in the word vector space. However, the Transformer Encoder can extract the global features of each text input, which is realised using a process of self-attention; thus, it can collect more text information than the first three extractors of text.

Baseline models to determine credit risk are taken from Machine Learning fields. Logistic Regression is one of the most common prediction models used in risk evaluation. The model assumes a relation between the explanatory variables and the default outcomes that can be expressed as follows: L(x) =1/(1+e^(wT x)) [1]. The random forest method is used for solving both classification and regression problems. The model randomly chooses k independent features from all available features and then uses these sub-features to build a decision tree. It independently repeats this process m times to obtain m decision tree classifiers. The resultant decision trees then vote on a label for each sample, and the predicted label is the one which receives the most votes [1]. Other models such as Support Vector Machine (SVM) and deep feed forward network which is based on artificial intelligence have also been used in credit risk modelling. P2P lending as a new and growing industry requires a more reliable model for assessing credit risk. One direct way to increase the quality of the loan assessment is to integrate more information into models to reduce asymmetry in information.

**[2] Use of Machine Learning Techniques in Financial Forecasting**

Financial institutions have a key role to play in speeding up economic growth and sustainability. In fact, one of their major sources of income is debts in the form of loans made available to their customers. But the cost of providing this service will lead to financial instability within these institutions. To prevent such events, there is a need for an appropriate credit approval process by various methods to forecast and evaluate delinquency and to prevent falsification. It is therefore essential to resolve credit risk management by creating a model that would identify potential borrowers as good or bad credits. The data is prepared by using historical collection of data with mixed attributes. Data preparation and analysis techniques such as univariate, bivariate using ANOVA and Chi-Square tests and K – Nearest Neighbour [2] have been used.

For a sample of 1000 individuals, the paper assesses the credit-worthiness of potential borrowers based on a group of attributes that define the characteristics of the lender that affect their ability to repay the loan and escape default and offence. In this paper, the success measure is the class of potential borrower (Good Credit / Bad Credit). To evaluate this classification, the K-Nearest Neighbour algorithm has been used by dividing the data set into a training data set and test data set. KNN's usage is related to the fact that, given the value of k, it is a simple but efficient classification method based on the selected neighbours’ majority vote[2].

Data collection was done using a premade credit card risk management dataset which contained 1000 observations with 20 attributes. This forms the basis of the classification. The data which has been provided has been made easier to visualize and understand by means of frequency tables and histograms which are used to identify the different classes and the distribution of the categorical variables. Central tendencies of the data are also measured to determine the distribution of numerical data. The research papers aim to achieve an accurate classification and for this purpose a bivariate analysis has been performed to Select attributes that have a significant relationship with the dependent variable and check if there is a correlation between the attributes. If so, either it is decided to remove one of them or create a new variable that sums up the information between the associated variables. Different significance tests are applied to achieve this purpose. Correlation Coefficient is a statistical significance test measuring the intensity of the association between the relative movements of two numerical variables. Chi Square Test signifies the measure of freedom determining whether or not two categorical variables are related. For one nominal variable the frequency of each category is compared to the categories of the second nominal variable. ANOVA Test measures the relation or significance between the dependent component, and the independent variable. It measures the variation of numerical variables among the different categories [2].

The model was trained, and it was used to get the precision and recall. The test sample could be classified with a high degree of accuracy using an optimal K value taken as 11 from the most significant variables. As the dataset contained both qualitative and quantitative data, the dataset was transformed from by turning categorical attributes into ordinal attributes so that the necessary algorithms could be applied to the dataset. Classifying prospective creditors into either good or poor accounts provides financial firms with an evaluation of their clients' creditworthiness. This is one potential mechanism to be used to avoid defaulted and delinquent accounts and to preserve the institution 's viability by growing its growth and profitability and therefore the economy's stability.

**[3] Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending**

Analysis of credit risk usually relies on, among others, statistical models such as logistic regression, probit regression, discriminant analysis, and Cox survival models[3]. These methods provide good efficiency, are easy to understand and do not pose problems in computing. On the other hand, alternatives to machine learning also provide better predictive efficiency, since they can detect more complex patterns of risk. Most methods of machine learning are usually black boxes, with little to no possibility of interpretation. As a result, banks are taking caution in implementing machine learning for modelling credit risk.

Machine learning model implementation requires complementary validation components, and, among other things, research and review of new prejudices or interpretability, even if they are not specifically comprehensive. Interpretability and accountability are important for various credit administration process models, such as processes for awarding, conduct and selection, or fraud detection. They are required to ensure a fair, controlled, and supervised credit delivery process. Regulatory agencies and end-users need interpretable models, and thus usually rely on models such as logistical regression. Machine learning algorithms are compared to logistic regression, which is a well-established credit risk methodology. A decision tree, which is a bagging classification method like random forest, and XGBoost, which is a gradient boosting classifier, is applied to the machine learning alternatives[3]. For each machine learning algorithm, we use genetic algorithms to search for an appropriate combination of hyperparameters, in the same line as other works. With regard to the comparison of results, we use statistical inferences to assess which method is doing better. For each class a detailed analysis of the classification metrics is provided. With regard to the comparison of explanability, we use the values of Shapley, as implemented in other such works. With the support of Local Interpretable Model-Agnostic Explanations and other methods, these works present a unified framework for interpreting predictions known as SHAP based on aggregations of Shapley values. The SHAP values enhance model transparency by offering global and local interpretability. Overall, they estimate how much each variable contributes to the target variable, either positively or negatively. The notion of SHAP values are extended to logistic regression. With the aid of graphical methods, the values for both approaches are calculated and evaluated. It is worth noting that the SHAP values are calibrated for categorical variables to better account for the interdependence of each group. The findings show that the approach to machine learning not only obtains better output outcomes but also sheds light on the complexities of the problem at hand, which is often overshadowed by linear approaches. More precisely, the SHAP values reveal that approaches to machine learning can detect complex nonlinear relationships that cannot be reflected by logistic regression, including dispersion and structural breaks[3]. Technological advancement has introduced new credit products like the peer-to - peer lending market. P2P loans act as loans between individuals, borrowers, and investors, linked by technological platforms. P2P lending lacks the intermediation of conventional banks, and thus the asymmetry of knowledge is much more pronounced than in traditional banking. Consequently, in P2P lending, calculating credit risk in an interpretable manner is much more difficult than with conventional goods[3].

**[4] Cluster Analysis for mixed data: An application to credit risk evaluation.**

Credit risk is one of the main risks facing a bank to offer financial products and services to customers. Many scoring methodologies have been developed to evaluate clients' financial results, and these methods are often focused on quantitative metrics. This paper highlights the importance of applicants' quantitative and qualitative features and introduces a new mixed methodology. Cluster analysis can prove especially useful for credit risk estimation. Clustering typically focuses only on quantitative or qualitative data at a time, but because credit applicants are characterized by mixed personal characteristics, a cluster analysis unique to mixed data may contribute to the discovery of especially insightful trends, estimating the risk associated with credit granting. Credit scoring is generally measured from the perspective of binary classification, sorting the possible risk into two groups: good and poor credit risk; Rejecting an applicant 's offer, which may repay the loan, will in one case constitute a business loss. In a second example, granting a claimant's credit request that cannot repay the loan would be a financial loss. The latter is the most detrimental scenario; indeed, the possibility that the debt will not be paid is clearly worse. Therefore, it is important that the bank assesses the risks associated with lending money to a client

Many clustering methods are limited exclusively to one type of variable at a time. For this purpose, it is common to turn mixed data into a single data type by converting the categorical variables into binary variables and thus applying methods for numerical variables or converting continuous variables into categorical variables. Less encountered in the literature are clustering approaches unique to mixed data. Clustering mixed data is a challenging job, since for these two types of data there is a significant difference between the similarity metrics. The main issue is the identification of a unified similarity metric for the sake of comparisons for categorical and numerical attributes.

Taking X as an (n x M) matrix to create clusters. The main aim is to obtain K clusters among n objects in X using a single cost function[4]. The methods of Ahmed Dey, Huang and Cheung & Jia are considered for mixed attribute comparisons[3]. The algorithm is known as the K-prototypes algorithms which is based on the K-Means method. The qualitative variables are useful for the proper detection of clusters, adding additional information for customer profiling. All mixed-data algorithms show a very small error in classification. The Huang method is however not without weaknesses. It assumes the weights of all numerical attributes to be 1 and the weight of categorical attributes as user defined. This has two main disadvantages, First, the numerical variables may not have the same relevance in a real-world dataset; further, in the case of incorrect user-defined weights for categorical attributes, an imprecise clustering could occur. Second, the prototype corresponds to the mode of the variables for the categorical variables. This leads to a loss of information, since only one attribute represents the cluster. Another issue is that for binary variables, the distance between two categorical attributes is not defined appropriately. Many of these issues have been fixed by the Ahmed and Dey method. Both internal and external validity indexes have been used to assess the clustering results.

**[5] AzureML Based Analysis and Prediction Loan Borrowers Creditworthy**

As conventual financial institutions offer higher-interest loans, the peer-to - peer approach to lending is more difficult. They charge fees, it remains cost effective for consumers because lending companies are more profitable. Lending club is one of those online platforms that match lenders with individuals. They have managed to attract more than 3 million clients, raising more than 50 million dollars, according to the lending club website. It has over 150 features, and over 2 million records. Through analysing this dataset, you can achieve monitoring and predicting individual credit behaviour. This can help banks to assist and approve loans request automatically with a lower rate of risk. A generic IoT used Azure machine learning framework created by Microsoft is used to ensure prediction accuracy. The reason, AzureML platform is used is because it has many good features that can compile the work since it's simple for developers to create, deploy, manage the application through Microsoft datacenter. Azure doesn't cost users a lot of coding time because it is drag and drop functions. It makes fast working for all sorts of datasets, such as excel, csv, tsv, etc. It makes it easy and fast to deal with all sorts of datasets. A generic criterion is introduced in this paper to analyse the dataset and understand the features, and to use them in the model. Then two Jungle Decision Classes and two Forest Decision Classes are used to compare results. The result is that the Two Jungle algorithm has the best results compared to the Two Tree algorithm.

The methodology first involves preparing the dataset. The Lending Club data set is used for this purpose. The dataset was then cleaned and updated; redundant data was removed for the dataset. Unnecessary and columns with unique values were removed because they weren’t important. Conflicts between the data was solved by ways of Data Integration. The machine learning algorithms used are 1. Two classes decision jungle 2. Two classes decision tree. Decision jungles are a recent extension of forest decision making. A decision jungle consists of an acyclic diagram (DAG) directed decision set. Decision jungles offer the following advantages: By allowing tree branches to combine, a DAG decision usually has a lower memory footprint and a better output in generalisation than a decision tree, albeit at the expense of much longer preparation. Decision jungles are non-parametric models which may describe boundaries of non-linear decisions. They perform integrated selection and classification of features, and are resilient when noisy features are present[5]. A decision tree is a predictive model that covariates the space of recursively partitions into subspaces, such that each subspace forms the basis for a different predictive feature. Decision trees may be used for different learning tasks like the study of grouping, regression, and survival. Because of their unique advantages, the decision trees have become one of data science's most strong and successful approaches. Decision forest aims to enhance the predictive efficiency of a single decision tree by practising and integrating its predictions with multiple trees[5]. The results of the paper show us that Two jungle is better than Two tree algorithm. Part of the objective was to use machine learning algorithms to predict whether borrowers will pay or not and as has been shown in the paper, the results of two jungle algorithms were good in terms of accuracy , recall and AUC[5].

**[6] Data Cleaning for Personal Credit Scoring by Utilizing Social Media Data: An Empirical Study**

With Fintech's booming success in the mobile network and rapid growth, Internet finance offers increasingly successful financial services for low-income groups as a valuable complement to the conventional financial industry. And the conventional banks aim to provide low-income citizens with small loans. Because of the lack of credit rating, credit risk is still the major source of risk for Internet finance and conventional banks in these financial services. Therefore, in view of this kind of financial services , it is important to create an effective framework of credit evaluation and to develop a model of credit evaluation to reduce this kind of credit risk. With the widespread use of social networking resources online, more and more human individual activities have been accurately tracked and a large social media dataset has been created, called "Big Data." These databases document real human interactions, which are part of real people's social mapping, and make it possible to use social media data to calculate credit level for users. The multi-source and heterogeneous characteristics of social media network data coexist with opportunities and challenges to use them for personal credit scoring as compared with structured data in traditional banking. To utilize such complex social media data, it is obvious that to deal with their complex structures and dimensions, more effective methods need to be created. Weibo data was used as the basis in this social media for extracting the age, tweets, and network features of user users. They used these three types of data as training data to create a structure for machine learning and by this they obtained an effective algorithm model for credit evaluation. In addition to Weibo data, cell phone data, shopping data and other data related to different aspects of individual everyday life, as well as various algorithms such as Logical Regression and Gradient Boosting are used for credit evaluation[6]. There seems to be a connection between conventional credit scores and data from the social network. These approaches affect conventional banks' data analysis. Myers and Forgy use SVM algorithm based on real credit card payment data, which demonstrates a very powerful credit evaluation capability.

With a credit score, it will determine the loans you are eligible for and the interest rate you are going to pay. The standard credit rating system is usually focused on individuals or organisations ' financial records. Inspired by conventional consumer credit ratings, a method of personal credit assessment based on online social media data has been developed as a method of credit evaluation for low-income groups. This approach is becoming a reality due to the advancement of the big data and machine learning model and algorithm. By evaluating potential abnormal users, we suggest three criteria for data cleaning, power exponents of individual user's time interval distribution, user's activity, and user's out- and in-degree ratio. We clean the star and net star users, "Water Army," marketing users, robotic users, etc.[6] systematically using these three criteria. The network structures established by the followers and followees were dramatically modified before and after data cleaning. The data cleaned up for personal credit scores are "Real but False Data". Based on data cleaning, we used the logistic regression approach to determine the individual credit scoring of users before and after data cleaning, and found that the order of personal credit scoring before and after data cleaning has changed considerably. This transition is largely due to improvements in network configuration following data cleanup. This thoroughly demonstrates the essential role of network structure in credit assessment using data from the social network.

**[7] Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model**

Credit scoring is a comprehensive statistical examination that lenders and other third parties perform to gain access to creditworthiness of a person. Lenders use credit scoring to measure a person's level of risk when lending money. However, the determination of credit score is based largely on a transaction record, payment history, professional background, etc. from various credit bureaus. Evaluating a credit score is therefore a laborious and tedious process which involves a lot of paperwork. A varied collection of data including the applicant's age , sex, loan intent, previously submitted applications, completed previous loans, work type, period in a work, type of accommodation, average bank account balance, and previous defaults are used to determine a reasonable credit score. All such data passes through several cycles of statistical analysis to produce a scorecard for a person. A reasonable credit scoring model is supposed to give the borrower a high score whose loans are expected to perform well and a low score to applicants who are more likely to become defaulting. The sophistication of the methodology involved in the actual credit risk assessment is rather repetitive. Various visual analytic systems have been created to make the correlations easier.

The process for credit scoring is accomplished by means of a blockchain which is a distributed ledger system. The proposed blockchain model is a cooperative or federated blockchain network where all the lenders and borrowers can access the network by providing their identity at any time. All transactions are encrypted using asymmetric key cryptography, in order to protect user data privacy. The user information can only be accessed and checked by intended borrowers using their private key. The users sign the transactions digitally using their private key to avoid the non-repudiation issues. All lenders will get a consensus credit scoring value[7]. The model takes the transactions, costs, and returns available to investors and/or lenders into consideration. It also guarantees accountability in all operations along with granting all persons privacy and anonymity. The model also takes into account the behavioural modelling of different lenders using Prospect Theory, based on their experience and different borrowers. A shared consensus forum to attain a unified credit score rating for all lenders. We also use the lenders behavioural model of lending money based on the credit score. The paper's main contributions are as follows: Decentralized & Unified assessment of the credit score Offering a decentralised and unified framework for assessing the credit score utilising blockchain. All transactions are registered, permanent and immutable. Behavioural Modelling for lenders: Based on decentralised credit scoring, we model the trade-off between risk and return on lenders using prospect theory. Reducing dependence on data sources: The availability of unchangeable public ledger eliminates dependence on various data sources and transaction information loss[7].

The credit rating is used to measure an individual's creditworthiness. It's a statistical parameter which quantifies a borrower 's risk of lending money. This paper presents a stable and decentralised model for calculating an individual's credit score removing the need for trusted parties and aggregating transaction history. The credit score measured in this model takes into account an individual's financial as well as non-financial information for determining the creditworthiness.

**[8] A STUDY ON CREDIT RISK MANAGEMENT IN SCHEDULED BANKS**

Credit risk management is a very critical field for the banking sector as there are broad growth opportunities and other financial institutions face financial problems. Banking practitioners often have to keep a balance between the risks and returns. Banks need a range of loan products to get a broad customer base. Credit risk is characterised as the probability of a borrower or counterparty failing to fulfil their obligations under the terms agreed upon. The more diverse a banking group is, the more robust the structures it will need to defend itself against a wide range of risks. These include the normal operating risks common to any business concern, the business risks to its commercial creditors, the economic and political risks associated with the countries in which it operates and the commercial and reputational risks associated with non-compliance with the increasingly stringent legislation and regulations relating to it. Credit risk management helps banks to proactively define, evaluate, manage and mitigate their credit risk at an individual level, at an organisation level or at a country level. Given the rapidly evolving, competitive world scenario under pressure from globalisation, liberalisation, consolidation, and disintermediation, it is critical that banks have robust credit risk management policies and procedures that are adaptive to these changes and react to them.

Plan for the analysis is adopted for exploratory studies. The collected data comes from both primary and secondary sources. The sample size is 30 percent. In this analysis the sampling approach is convenience sampling. Questionnaire is used in collecting data. The questionnaires are individually administered and also mailed to the respondents and circulated electronically by email.

SUGGESTIONS AND RECOMMENDATIONS

The bankers will use compliance by getting village heads socially valued individuals to validate the borrower 's reputation.

Training bank managers to recognise and eliminate bad debts as to be extended. That will provide the bank with a long-term profit.

Amend policy on credit. Some items can be tweaked or changed which will make it harder for consumers to get accepted when the economy is experiencing a recession. During more prosperous times the strategy can often be reversed. A lending institution may want to remain at the verge of going over what is a reasonable amount of debt away from customers. If a consumer needs capital for home improvements, then debt in the process can be considered. Collateral demands can also entice a consumer to make payments.

Reduce credit limits for consumers at high risk. Many creditors will slash your credit cap if they consider like you are a high-risk client. Your credit limit will be reduced when payments are late or if you have too much other debt. This prevents you from making additional purchases when layoffs are a possibility for many industries particularly during difficult economic times.

Credit Risk Management research has shed light on many fascinating issues of Scheduled Banks. The aim of credit risk management is to optimise the risk-adjusted rate of return of a bank by holding exposure to credit risk within appropriate parameters.

Banks are increasingly facing credit risk (or counterparty risk) in a range of financial instruments other than loans, including acceptances, inter-bank transactions, trade finance, international exhcnage transactions, financial futures, derivatives, shares, securities, options and the extension of commitments and guarantees and transaction settlement. Efficient credit risk management is a critical component of a holistic risk management approach, and is important to any banking organisation's long-term success.

[9]**CREDIT RISK MANAGEMENT IN COMMERCIAL BANKS**

In order to ensure successful credit risk management in commercial banks, it is important to establish the kinds of terms and conditions that will attract potential borrowers and guarantee loan repayment for those bank clients who take out loans. However it would not be expedient for each individual borrower to establish a separate set of terms and conditions. Instead current and future bank customers should be classified by their similarities and difference. After that, it is appropriate to work out a different set of terms and conditions for each group according to the characteristic features of the group members. The classification of bank clients into separate groups should proceed according to the classification method which unites disparate elements of the system into homogeneous groups based on the similarities of the elements in question. This classification method must reflect the structure of the source data and ensure that the data are divided into groups in the most appropriate way. Clustering and networking have historically been employed to attain these objectives. Both of these approaches in the case of multidimensional samples generate identical divisions of objects into groups. We will use clustering as the method of credit risk assessment in this article. To determine the risk of lending activities of a bank, it is important to take into account the statistics representing breaches of contract terms by bank customers and the harm caused to the bank by each such breach. The extent of the risk as the amount of harm (risk defined as failure of the borrower to make principal payments on time) can be considered as being aggressive dependence on factors such as the average size of the loan 1 x, the duration for which the loan is given 2 x, and a variety of other factors. Specification and detection of such regressions should be focused on the knowledge about the harm incurred by each customer and the credit characteristics of each customer group. Such a model will allow any potential client to predict the risk posed. In managing credit risk, a framework of interconnected and interdependent methods of deliberate action needs to be developed with a view to reducing risk and uncertainty in crediting activities. Using the proposed credit risk assessment model helps the credit risk management tential client to follow a differentiated approach. Credit risk management can be interpreted as a process consisting of the following phases : 1) identification of the risk factor; 2) evaluation of the potential consequences of the risk factor identified; 3) choice of management strategies to mitigate the consequences of the risk factor identified; 4) supervision (monitoring) of the implementation of the chosen risk factor;

The potential risk is evaluated at the credit risk identification stage in terms of its quantitative and qualitative criteria in the sense of the bank's risk factor analysis to evaluate the degree of severity posed by the risk concerned. At the identification stage of a possible credit risk, the effects of managing the identified risk can also be expected in consequence of the different sets of management methods used; In order to be able to select the best set of methods to be used in the future according to the criteria identified above, one is thus allowed to compare and contrast different set of risk management methods in order. Often, from the perspective of the extent of their impact and probability of occurrence, it is necessary to assess the implications of a potential credit risk. The credit risks defined at the first stage of credit risk management must be measured on the basis of the following temporal parameters: past data, current data and forecasted future data. Given that credit risk factors appear to change over time , it is important that these changes be monitored in order for bank management to have a timely response to the credit risk rise relative to the expected credit risk value. Credit risk control is the core of credit risk management and the aim of the final level. What needs to be underlined is the functional and organisational consistency of all the credit risk management levels. Indeed, all phases of credit risk management are inextricably intertwined, and the core concept of credit risk management is their unification. The differentiation of methods used in credit risk management in different groups of borrowers is due to the Basel Committee 's international standards and requirements and will contribute to the transition of banks to the use of an internal rating approach when assessing and managing credit risk. This limitation determines the research's future direction, which will involve studying credit risk management in those cases where loans are given to other categories of borrowers, such as small and medium-sized enterprises ( SMEs) and large industrial enterprises.

**[10] Impact of risk management strategies onthe credit risk faced by commercial banks of Balochistan**

In 2008, the credit crisis began worldwide as a result of mass issuance of sub-prime mortgages to individuals in the U.S. leading to defaults, which caused financial institutions worldwide to have outwardly-rippling problems. Sub-primary mortgages and other loans with lower restrictions can generate significant losses for financial institution including corporate failure and bankruptcy (Brown & Moles,2014). These credit decisions play a pivotal role in profitability for companies. While the decision to over-extend credit to high-risk customers may increase short-term profitability for individual banks, this lending behaviour has been seen as a major challenge for the economy as a whole's risk management structures. Managing risk therefore is the most important element in the operations of a bank. This phenomenon is equally applicable to global banks, including Pakistani banks. Because of the unstable and volatile nature of the political and financial environment in Pakistan, many types of risk affect banks including foreign exchange risks, liquidity, operations, credit and interest rates. Pakistan's financial institutions are generally risk-averse, particularly with regard to car finance and mortgage loans, with higher chances of huge losses (Shafiq & Nasr,2010). Balochistan is the most underdeveloped part of Pakistan with the largest geographical area. Small businesses have limited opportunities and the majority of businesses are run with poor documentation in informal form. Most commercial banks face problems like the verification of loan documents and the processing of loans. Appropriate risk management strategies can therefore help to understand and mitigate the credit risk faced by Balochistan's commercial banks. This study aims to identify the various strategies for risk management that may influence credit risk management by commercial banks. We expect to determine whether these strategies will contribute to both credit risk reduction and effective performance in meeting customer requirements.

The first hypothesis considers assessing the role of hedging in reducing the credit of a bank. Based on a model presented by Felix (2008), which showed hedging risk management strategies, capital adequacy ratio and diversification can be used to explain a bank's credit risk. So our first hypothesis is:

H1: hedging will minimise credit risk faced by the Balochistan commercial banks.

The second risk management strategy is diversification, which requires banks to provide a wide range of financial services with flexible terms to customers and to providecredit to a wide range of customers instead of few in order to reduce risk (Fredrick,2013). Banks can use the concept of diversification as they create a pool of widecustomers to provide loans, rather than providing large amounts of loans to few customers, which inherently increases the risk (Hobson, 1998).

H2: diversification will thus minimise credit risk to Balochistan's commercial banks.

The third hypothesis considers management strategy requiring banks to maintain the capital's aparticular amount (Ho & Yusoff, 2009). The capital adequacy ratio is critical for banks to be in a better position to manage unexpected risks and therefore capital maintained in a bank has an impact on the overall credit risk, so it can be assumed as follows:

H3: capital adequacy ratio will minimise credit risk of Balochistan commercial banks.

The fourth hypothesis takes into account the role Corporate Governance plays in minimising credit risk. Corporate governance assumes that the organisation or corporation should adopt all practises that ensure accountability to stakeholders (Shafiq & Nasr, 2010 ) . Consequently,

H4: corporate governance will minimise credit risk to Balochistan's commercial banks.

Several banks have failed in the past because they were unable to control their credit risk. Recommendations for banks resulting from this study include the diversification of their products and services, which is critical as it allows the bank to provide many products and services to customers. According to the findings, an emphasis on employing corporate governance policies after diversification is most important. Hedging and the ratio of capital adequacy are

also important strategies that banks can examine and opt-out. Hedging is useful, because flexible contracting helps to reduce the risk. Balochistan's banks will be able to realise the importance of the capital ad-equacy ratio as this will allow them to strike a proper balance between the amounts of capital that should be maintained in order to manage investor needs. It is recommended that further research on the topic be carried out so that banks can identify effective strategies for managing other risks. These banks 'success and further progress depends on the smooth implementation of risk management strategies and activities, which have been demonstrated to have a very significant positive impact on the Balochistan banks' ability to control credit risk.

**[11] The Relationship Between Credit Risk Management Practices And Profitability In Malaysian Commercial Bank`S**

Credit risk management is the method of removing all possible risk factors affecting investments of any kind. The last global financial meltdown underscored how important credit management is to the finance sector. Not to mention how it is increasingly becoming a matter of concern which involves ingenious methods of risk management. Hence the main purpose of this paper is to explore the degree to which current practises in credit risk management on profitability of commercial banks are successful. This research was designed and formulated

using a system of quantitative research design. We have used cross-sectional survey techniques to gather all necessary data. Banks usually face several challenges, the main of them being credit risk and also considering that credit is the number one source of revenue for most banks. The harm or failure may surpass all target goals without the right strategies being put in place. After reviewing a wealth of field studies, we find a lack of data that describes consistently the relationship between profitability and credit risk on Malaysian commercial banks. Managers of commercial banks should deepen their knowledge of credit risk management through successful preparation. And continuous learning about the latest developments and strategies, enhancing their banks' overall performance and minimising any credit risk opportunities. Nowadays, the main risk faced by most banks is credit risk and can be described as: the possibility of borrower failing to fulfil their obligations (primary, interests, and commissions), on time or in compliance with the terms agreed upon. Banks are required to build Loan Loss Reserves by statute. To offset the losses resulting from the loans. Credit risk comes from the unlikely payment by a debtor of his obligations

or the weakening of his financial ability resulting in economic loss for the bank. Risk management 's key goal is to minimise the danger from benefit fluctuations and to prevent or reduce the effects of any major losses. Normal risk management techniques are to define risk areas, to calculate the magnitude of impact. And, Apply the right tactics accordingly. It all starts by first inspecting and recognising the sources of any possible risks or future threats and then beginning the risk management process. Second, the risk found during the detection process must be quantified by the calculation. For example, a person needs to calculate the likelihood of the actual default and consider to what degree the effects of the risk may affect the default probability. At this point, statistical analysis for the accurate calculation of the risk must be carried out. In January 2018, the Central Bank of Malaysia (BNM) finalised the credit risk analysis standard to replace the 2001 protocol. Examination criteria improve various facets of credit risk management activities at financial institution. These criteria include a range of items, rising financial institutions' cross-border business and further expanding the domestic capital market, which is an alternative source of funding for large banking institutions. The revised standard allows financial institutions to follow a more sophisticated method of calculating credit losses and also supports Malaysia Financial Reporting Standard 9: Financial instruments (MFRS 9), which came into force on 1 January 2018. Forward-looking approach Financial institutions are expected to predict potential credit losses across the credit facility, taking into account wider macroeconomic trends, according to new impairment requirements from MFRS 9. While higher provisions are expected to apply to MFRS 9, the effect on bank earnings and capital is projected to be within bank expectations. A research paper without a proper methodology is totally worthless and far from having any effect. It sets all the requisite principles for a researcher to carry out his work, from data collection to data analysis. The researcher chose to employ a quantitative approach, which is precisely subtle for the purpose of this paper, to clearly define the connexion between profitability and credit risk. There are a number of research projects clearly designed to solve market problems. Research designs are triple, depending on how research formulates its methods and aims in a specific article. They are explanatory, descriptive and exploratory study. To obtain the best results, a researcher should select a population with metrics that serve the function of his research such as particular class of persons or events. A collection of all factors deduced from population samples is called a sampling frame, and is therefore the precursor to any successful survey. It is a list of all local commercial banks with a population which can be sampled in Malaysia. The target population of this study has been the Malaysian local commercial banks. Therefore, this study's target audience which are the commercial banks that contribute to the phase of economic growth. The primary aim of this analysis is to use financial ratios to measure and dig into commercial activities of all banks. The total number of banks considered tradable in Malaysia is 25. These are divided into eight local banks and seventeen international banks. But this is a survey study of all commercial banks as they make up the target population of this inquiry. As of 2017, there are 8 fully operational commercial banks, all of which are approved and locally based. Ultimately, the meaning and intent of this paper has been successfully achieved by reviewing and studying the various effects on profitability among local banks in Malaysia of both dependent and independent credit risk variables.

**[12] What’s wrong with modern credit risk management?**

Bank security against credit risk implies an implied secret insurance contract. To resolve this shortcoming an approach has been developed. "An allowance for expected credit losses should cover fully realised losses within expected losses" The capital of a bank should cover a portion of credit losses that exceed expected losses. It is seen as the effect of fiscal and monetary policies on the economy's structural level of defaults. The credit loss provision is aggregate, impersonal (not connected to a particular borrower) and expected. As concrete borrowers go bankrupt, the individuality of the estimated credit losses comes out. So credit risk gets a profile, so to speak. It becomes clear to the creditors the credit losses actually belong. From a mathematical perspective, a bank should assume that out of a small number of default borrowers, an amount generated by payments from all borrowers would cover its actual credit losses. The bank uses only a small portion of the fund to compensate its losses sustained. It protects most of them by unnecessary extra expenses. The bank will completely cover its expected losses through the already-formed allowance. In our case, the bank's allowance nearly doubled. The need for an additional increase in the bank's allowance due to borrowers' defaults creates the impression that this payment is inadequate to cover credit losses if these are either less or equal to the anticipated ones. So surprisingly, the proclaimed "Allowance covers expected credit losses" principle does not work. Here there is something incorrect, namely the method for using an allowance in case of default. To restore this idea, we suggest the procedure that follows. In the beginning, we will differentiate between two types of allowances for banks which are for planned and actual credit losses. An allowance for expected losses is collective, whereas allowances for realizedlosses are private, linked to particular default borrowers. In the case of default, private allowances for realizedlosses are generated from collective allowances in the maximum unpaid sum of bad loans. That is, making provisions for anticipated credit losses would not entail any extra expenditure from the bank. Only then can properly use the allowance for expected credit losses.

We assume that all financial market participants, including banks, and insurance firms should operate according to the same guidelines. Then, in order to improve banking lending transparency, it should be understood that there is a secret, tacit credit risk insurance policy. Those contracts ought to be made public. For this, a credit risk premium and, most significantly, the amount of insurance coverage for and borrower need to be clearly stipulated. In other words, when a bank signs a credit contract with a borrower, it will assess and notify the borrower about the amount of insurance and the credit risk premium. The risk premiums are charged by all borrowers to form a mutual reserve for potential credit losses. The loan provides for fulfilling the responsibilities of the borrowers towards a bank. But if any creditors go bankrupt their loans due to mutual interest will be repaid in full.

If a borrower owns assets that can be pledged, then three options are available:

- Pay a risk premium, receive a higher interest rate, and pledge no collateral. In this situation, the perceived credit losses should be entirely compensated by a mutual allowance;

- To pledge collateral and pay no risk premium, correspondingly, to get the lower prime interest rate. In this case, the collateral should compensate the credit losses realised in full;

- To combine the latter two options when a collateral value is insufficient to cover the credit losses realised in full.

Both the government and the central bank should be active in offsetting the credit losses incurred by their fiscal and

monetary policies. "Once NPL issues become widespread, a concerted and structured solution becomes important and is usually led by authorities. Unable to address system-wide NPL situations through a bank-by-bank approach. It is proposed to differentiate between two types of provisions for banks which are for credit losses predicted and realised. A payment is collective for expected credit losses, whereas allowances for actual credit losses are individual, linked to particular default borrower. The established approach suggests that the credit risk premium allowance should compensate the losses suffered by default borrowers. In this case, a bank does not bear any additional charges for the provisioning of allowance.

The strategy has the advantages of:

· A reduction in spending on the payment of loans (payment for credit losses in the bank's income statement);

· More reliable and direct monitoring of both planned and unforeseen credit losses;

· This results in a decline in bank defaults.

The developed approach also suggests recognising hidden, embedded in lending, and making it explicit, a credit risk insurance. For this, it is crucial that banks stipulate explicitly the amounts of credit risk premium and, most significantly, the amount of insurance reimbursement. Insurance reimbursement from allowance would minimise the burden on indefault borrowers. Each borrower has the right to such a reduction because it has entered into insurance contract with bank. Moreover, credit transparency will improve for the borrowers.

**[13] Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer**

In this study, we used deep neural network model to classify data from the UCI machine learning repository called the default payment by Taiwan credit card. Wetriedto covereverystepand every concept needed to create a complete model. The data set is applied to financial areas that require high risk and use deep neural networks to reduce those risks. So, we used the default payment credit card dataset to predict the correct class of default customer by using step-by-step approach. The hypothesis for this research is "How can we achieve high precision using artificial deep neural network by classifying default credit card customers?”. The rest of the paper structure is as follows: Section II provides a written analysis or examination of the literature. Section III describes the various relevant words with the aid of previous research. Section VI sets out methods and techniques for our deep neural network work.

Section V explains the exploratory setup used to introduce the proposed approach and its implications for default client classification. Section VI deals with debate, and Section VII addresses the study's conclusion and possible scope.

[1] This paper discusses the classification and neural network very well. The author outlines the generalisation of the neural network model as well as describing other methods of classification but this paper includes only theoretical concepts.

[2] This article discusses the default credit card payment data collection using Rapid Miner's data mining and machine learning platform.

[3] This article discusses the Taiwan credit card default payment dataset with six methods of classification and contrasts the effects of each approach with each other. The author concluded that the neural artificial network is working well.

[4] This article discusses the backpropagation algorithm and the neural artificial network. According to this article propagation back is a three-tiered architecture while the artificial neural network is the best approach to the problem of classification.

[5] This article provides for insolvency protection by means of artificial neural networks. In addition, it elaborates various terminology associated with artificial neural network. The author explains the tuning parameters of artificial neural network.

Classification is a popular method of decision-making about human activities. Classification is a technique that uses to group-objects based on their characteristics in various classes. It addresses discrete data and predicts discrete or categorical categories of classes. Classification is defined as a technique where we classify data into a certain number of classes and want to assign a category or class to which new data belongs or may fall within. Classification model is used to forecast unseen or test data for new class category. Classification may be in three ways such as binary, multi-classification, multi-label classification. There are three different methods for classifying the very first statistical model: naive Bayes, support vector machine, k-nearest neighbour (non-parametric) second, decision tree (rule-based models), random forest, and the third method, known as artificial neural network, is now the advance form of the neural network. It is the most commonly used method of classification that improves the accuracy of the model. Deep neural network is the most common of all classification techniques in the areas of finance and health. Previous research shows also that deep neural networks perform better than any other system of classification.

In this work we clarified the deep neural network model and the default customer dataset for Taiwan credit card. Previous work does not clarify the deep neural network model and the problem of overfitting in this way when training the model. We tried to clarify the various methods of regularisation that can be used to generalise a stronger and more consistent model. It also incorporates a full step-by - step approach. We have shown that using deep neural network, we can identify default credit card customers and get the accuracy about 82-83. By following the process, this research prevails, we can make a detailed model of deep learning decision making. A day now the deep-neural network is very famous. It's being applied for prediction and classification in various fields.In this paper we defined deep neural network (DNN) a step-by-step method for the complete and accurate classification of customers default payment. We performed better than previous studies. We mentioned all steps that are needed with the aid of the Taiwan credit card default customer dataset to create the best deep neural network model and proved our hypothesis very true.

**[14] P. Giudici and L. Parisi, “CoRisk: Credit risk contagion with correlation network models,” *Risks*, vol. 6, no. 3, pp. 1–19, 2018, doi: 10.3390/risks6030095.**

We model CDS spreads using a structural vector autoregressive model, composed of a country-specific component that depends on time, and a contemporary component that describes the effects of contagion between countries. The proposed model refers to ten countries that are indicative of the recent financial crisis: top borrowing / lending countries, and European peripherals. Finally, the results of the study suggest that core countries are risk importers, as contagion increases the spread of their CDS while peripheral countries are risk exporters.

Without loss of generality, this paper will concentrate on sovereign countries with the goal of capturing the effect of contagion risk on their credit risk and, by doing so, on the flows of capital to and from each region, the variations of which can have a significant impact on financial stability. From an economic point of view, when a nation has a negative relationship with distressed countries, its final default likelihood decreases because it is viewed as a flight-to-quality refuge, and is thus positively influenced by the impact of contagion. Proposed methodology Let it be, at the time t, the Credit Default Swap spread of a country I for I and t. We assume that yit depends on an autoregressive component expressing the dependence on the past CDS spread values of the same country; a cross-sectional component expressing the contemporary dependence on other countries' spreads; a stochastic residual. Formal, for each country I and time t, we assume that the following parameters are: yit = p0 αip yit− p βij yt eti, p = 1 j 6 = I Features 2018, 6, 95 5 of 19 where p is a temporary lag, αip and βij are unknown parameters, and eti are standard Gaussian residuals, assumed to be independent over time and nation. Equation models CDS spreads through a VAR mechanism whereby the sovereign risk of each country p depends on its past values through the idiosyncratic component p0=1 αip yit− p, and the values of the other countries through the systemic component j6=i βij yt, which we term 'Contemporary Risk.' Equation justifies the substitution of βij and ji by their corresponding partial coefficient of correlation πij .. The partial correlation coefficient, from an economic point of view, expresses how the CDS spread of a country I is influenced by the contemporary spreads of the other countries j 6 = I The worse the countries I'm more linked to, the greater the predicted likelihood of I itself.

A country's default likelihood at time t can be either greater or lower than its default likelihood at time t1, depending on the sign of partial correlation coefficients with other countries. If λij > 0, the default country I likelihood increases after j contagion has been included. In comparison, if πij < 0, the default country I likelihood decreases after contagion has been obtained from j. Another benefit is the option of using network correlation models, where the matrix for correlation is represented. Negative contagion can be explained through capital flows: when a country i is facing a crisis period, investors tend to shift their portfolio towards "Safer" places in order to reduce the level of risk: such places are typically the countries negatively related to i, thus implying an improvement in their survival probability.

In more detail, the CoRisk values reported in Table 3 indicate that Greece is the country mostly impacted by contagion, followed by Ireland, France, Germany and, to a lesser extent, the United States and the United Kingdom: all these countries report a positive CoRisk value, indicating they are "Importers" of risk.

Such contagion effect can be both positive or negative: in the former case, it determines the spread to increase, typical behaviour of "Risk-importer" countries; in the latter situation, it causes the spread to decrease, typical behaviour of "Risk-exporter" countries. Negative contagion can be explained by capital flows: when a country faces a time of crisis, investors appear to move their portfolio to "Safer" places to reduce the level of risk: these places are usually negatively linked to I indicating an increase in their likelihood of survival.

More precisely, the CoRisk values recorded in Table 3 show that Greece is the country most affected by contagion, followed by Ireland, France , Germany and, to a lesser extent, the United States and the United Kingdom: all of these countries report a positive CoRisk value, implying that they are "importers" of risk. Such contagion effect may be either positive or negative: in the former case it decides the widespread, typical behaviour of 'Risk-Importer' countries; in the latter situation it causes the widespread, typical behaviour of 'Risk-Export' countries to decrease.

**[15] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, “Credit risk prediction in an imbalanced social lending environment FinanceIT Research Group , University of New South Wales , Centre for Artificial Intelligence , University of Technology Sydney ,” *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, p. 925, 2018, doi: 10.2991/ijcis.11.1.70.**

Au Abstract Credit Risk Prediction is an effective way to assess whether a potential borrower can repay a loan, particularly in peer-to - peer loans where problems with class imbalances are prevalent. Few credit risk prediction models for social lending consider imbalanced data and, in addition, it is still controversial to use the best resampling technique for imbalance data.

The findings show that combining random forest with random under-sampling can be an efficient technique for estimating the credit risk associated with loan applicants in the social lending markets.

Introduction Social lending, also known as peer-to - peer lending, uses online trading sites as a medium to lend money without conventional financial intermediaries, including banks, interfering. Furthermore, P2P lending typically takes place in settings with a high degree of knowledge asymmetry-that is, settings where the lenders do not have full details about the credit history of the borrowers. For these lending platforms to increase their market share in the financial industry, more powerful loan appraisal tools are needed[2, 4]. Traditional methods of loan assessment presume a balanced distribution of misclassifications, but an imbalanced dataset is far more representative of social lending platforms.

First, we are designing a new credit risk prediction model based on analytical intelligence approaches, and apply the current lending club dataset, one of the largest online lending platforms for P2P. Section 2 includes a literature analysis of the techniques of loan assessment in P2P lending markets and study related to class inequality issues. Loan Appraisal has appeared in the financial marketplace as a new e-commerce site in P2P lending. Research Methodology To help lenders evaluate the creditworthiness of borrowers in social lending platforms, we developed a decision support system that includes a novel prediction model to reduce the risk of loan defaults.Dataset Description Advances in P2P lending markets have generated large amounts of data on real-world P2P lending transactions.

Lending Club attributes loan characteristic Target variable Category = Where DTI, annual income, and installment amount are Lending Club features.The Lending Club assigns an LC grade of between A and G, where A represents safer loans and G represents riskier loans. To calculate the creditworthiness of borrowers in P2P lending platforms, we used the most recent data published by the Lending Club.

**[16] P. Monnin, “Integrating Climate Risks into Credit Risk Assessment - Current Methodologies and the Case of Central Banks Corporate Bond Purchases,” *SSRN Electron. J.*, no. December, 2019, doi: 10.2139/ssrn.3350918.**

If central banks underestimate risks by overlooking credit risk related to the environment, they could infringe this rule by accepting assets as collateral that do not meet these stringent risk standards. If credit prices do not accurately reflect climate credit risk, they send out signals and 1 See also Mathiesen for a discussion about how credit agencies are likely to underestimate climate risks, or Monnin for a more general discussion about why financial markets do not completely reflect climate risks. To illustrate how incorporating climate credit risk could change the operations of central banks, Section 4 applies one of the methodologies available for assessing the transition risk in the European corporate bond holdings.

To better understand why, we first introduce the notion of climate risks and the various types of risks it involves, and then explain how those risks translate into credit risk. WHAT Threats ARE CLIMATE? The NGFS, as well as the G20 Sustainable Green Finance Study Group, describe two key sources of climate change-related financial risks: physical risks and transition risks. Transition risks Transition risks can be described as the risks associated with the process of transitioning to a low-carbon economy of economic dislocation and financial losses.

HOW DO TRANSLAT INTO CREDIT RISK CLIMATE RISKS? Credit risk is the danger of a financial loss arising from the inability of a borrower to pay back part or all of the interests and the principal of a loan.

METHODOLOGICAL CHALLENGES Tackling five challenges is common to all methodologies estimating the credit risk effect of climate change, in terms of both physical and transitional risks: 1) resolving the limitations of historical data, 2) extending the scope of credit risk models, 3) seeking the right degree of data granularity, 4) defining the appropriate climate risk exposure indicators, and 5) translating.

One big explanation why market participants don't even perceive climate risks is that climate change related threats are believed to lie behind their conventional one to three-year review horizon. We see two reasons why using a shorter timeframe is not satisfactory: first, because "while financial risks can be completely realised over an extended time horizon, the risks call for short-term action to mitigate long-term impacts," and second, because "It is naive to assume that when risks become perceptible, everyone will be able to reduce their exposures simultaneously and in an orderly fashion. Translating economic impacts into financial risk indicators A final task is to convert economic impact projections into measurable and practical credit risk indicators such as the probability of default, default loss and thus credit scores.

Consequently, accounting for climate risks in the credit risk assessment of central banks, either by including them in their internal credit risk systems or by adding them to the credit ratings of credit agencies, is likely to change the collection of assets that they currently accept as collateral or purchase: once climate risks are properly accounted for such securities may not meet the risk requirements developed for such operations. To implement risk standards-including those at central banks-it is therefore important to incorporate climate risks into credit risk analysis.

**[17] Y. Cui, S. Geobey, O. Weber, and H. Lin, “The impact of green lending on credit risk in China,” *Sustain.*, vol. 10, no. 6, pp. 1–16, 2018, doi: 10.3390/su10062008.**

Based on a five-year dataset of 24 Chinese banks, we used panel regression techniques including two-stage, least square regression analysis and random effect panel regression to analyse whether a higher green credit ratio decreases the non-performing loan ratio of a bank. These regulations are compulsory for all Chinese banks, regardless of ownership structure, and thus include banks, joint-stock banks, and credit unions controlled by governments.

Together with 29 banks, a Green Finance Committee has been set up to coordinate activities such as establishing a green bond standard, promoting environmental stress tests for the banking sector, and organising discussions on greening China's overseas investment. While all of these studies indicate that green lending raises banks' risk of credit exposure, they do not address institutionally quite the distinct Chinese banking system. Exogenous Variables Credit quality reflects the credit quality of a bank: Credit quality CQit= Loan loss provisionit Total loan Sometimes, high NPL ratios are seen in low ROA banks. Banks with a lower ROA could take higher credit risks to improve their return.

Sample as an indicator for the size of a bank: Bank size Equation: Bank size Total Total assets assetsitit SIZE SIZEitit== 24 24 Total assetsit ∑i𝑖𝑖=1 Total assets it =1 The endogenous, andand instrumental variables that are used the model aremodel presented endogenous,exogenous, exogenous, instrumental variables that areinused in the are in Figure 1

Instrumental variable Type of bank Endogenous variable Green credit to total loans ratio Credit Quality ROA Exogenous variables NPL ratio Inefficiency Solvency ratio Bank size Figure 1.

First, the the banks banks that that are are considered considered Banks "Major banks" banks" in in China China were were included. According to the CBRC, there are 21 "Big Banks," including all policy banks, state-owned commercial banks, banks, and national joint-stock commercial banks, banks, as well as the Postal Postal Savings Bank of Trade China.

Included in the original bank list were three Originally, policy banks, five state-owned commercial banks, twelve national joint-stock banks, one postal savings policy bank, five state-owned commercial banks, twelve national joint-stock banks, one postal savings bank, 14 city commercial banks, and four rural banks.

Due Bank, Business Banks, Rural Commercial Banks. We analysed a sample of 24 banks, together with the exclusion of China's Agricultural Development Bank and Guiyang Bank due to lack of financial data.

Banks with significant volumes of state-controlled shares and state-owned businesses, such as policy banks and state-owned commercial banks, are more likely to grant a large proportion of green credit based on the two-stage model performance. The major impact of credit quality, ROA, and inefficiency, as well as a bank's size indicates that Chinese banks' NPL ratio is affected by factors that also affect banks outside China's NPL ratio.

**[18] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.**

Six approaches have been retained: a random forest model, a gradient boosting machine and four deep learning models; For the parameters of each model we specify the models and the choice retained. Describe algorithms for executing similar Lasso models: they offer fast algorithms to fit generalised linear models with the elastic net penalty. For many models the previous approach works: Linear regression: the answer belongs to R. So we're using that model.

The Models The models we are using were outlined in the section above. Output of models on the validation dataset with 181 variables using the values AUC and RMSE for the seven models. Test data set output of models with 181 variables using AUC and RMSE values for the seven models.

Models AUC RMSE M1 M2 D1 D2 D3 D4 0.876280 0.993066 0.994803 0.994803 0.904914 0.841172 0.975266 0.897737 0.245231 0.096683 0.044277 0.114487 0.116625 0.116625 0.113269 On the validation set, we observe that the AUC criteria are of the highest value with model M3, then model M2, then model D3, and model D1, then model D4, M1 and D2 for the last four places. If we consider the RMSE criterion, the M3 model gives the smallest defect, then the M2, then the D1 and D4 model and finally the D2, M1 and D3 model. Therefore, the largest error is in model D3.

If the gradient boosting model remains the best match for the smallest error, we can observe model stability on both the validation and test sets. We find in all cases that the deep learning models don't outperform the tree-based models. With regard to the AUC and RMSE, the logistic regression model and the multilayer neural network models (deep learning D1-D4 considered in this study in both the validation and test datasets using all 181 features, we observe that the gradient boosting model showed high performance compared to the random forest model for the binary classification problem, given the lower RMSE values.

Model M2 selects three variables from model M1 already given. Model M3 selects only one element that M1 includes. Product D1 uses three modell M1 variables. The D2 model selects one variable that is selected by M2. Model D3 chooses one variable used by model D1. Model D4 chooses one variable chosen by M1. The description of the variables used by each model is as follows: flow data and aggregated balance sheets correspond to the variables A1, ..., A10 of model M1. They have shown that, given the underlying model, selection of the top 10 variables, based on the variable value of models, does not inherently produce stable results. Our strategy of re-checking model performance on these top variables can help data experts validate the choice of models and variables selected.

**[19] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, “An Interpretable Model with Globally Consistent Explanations for Credit Risk,” pp. 1–10, 2018, [Online]. Available:** [**http://arxiv.org/abs/1811.12615**](http://arxiv.org/abs/1811.12615)**.**

Rather than introducing and describing a black box model afterwards, we give a globally interpretable model that is just as accurate as other neural networks. A two-layer additive risk model is what we term our globally interpretable model. Traditional subscale models are generally interpretable because they are decomposable into relevant components, and because these models are usually linear with coefficients whose sign for risk factors is positive. The question posed above, about a black box model being explainable, versus having a globally interpretable model, is significant. Second, whether the global model is consistently equivalent to the local model, often the explanations may be inaccurate, making it impossible to trust either the explanations, or the global model itself.

Even if an explanation provides the same prediction as the global model, it may be inconsistent with the actual estimates of the global model, or it may provide explanations which may be valid in some cases but not others. While our global model can be interpreted and can therefore be clarified on its own, we can also generate optional local "Explanations," which are now simply summarising the global model's general trends. The novel elements of the work are the two-layer additive risk model that naturally lends sparsity, decomposibility, visualisation, case-based reasoning, function value and monotonicity constraints, the model's interactive visualisation method, the use of the SetCoverExplanation algorithm for high-support local conjunctive explanations and the funding application.

First, we used step functions as our initial transformations of the features to ensure the model's monotonicity with respect to any given features, and restricted the coefficients of step functions of each feature to be non-negative. -Each subscale can be interpreted as a miniature model for predicting the likelihood that a loan may fail to repay, using only the features specified for the subscale. 3 Consistent rule-based explanations with SetCoverExplanation As noted earlier, apart from predicting risk using a globally interpretable model, we generate consistent rules that summarise the classifier's large data patterns.

**Modules**

* Data Preprocessing
* Missing Values Treatment
* Frequency Analysis
* Data Exploratory Analysis
* T-Test
* Visualizations
* P-value based Selection
* Multicollinearity with VIF checking
* Model Build and Model Analysis
* Train and Test (Logistic Regression)
* GINI rank ordering
* Accuracy
* Decile Analysis

**Proposed Algorithm and Implementation**

**1. IMPORT necessary python libraries for machine learning and data visualization (graph plotting) such as sklearn, numpy, seaborn etc.**

Every python program starts with importing the necessary python libraries. For machine learning modelling, data visualization etc, we require libraries such as pandas, sklearn, numpy, seaborn, patsy, scipy.stats, pyplot, matplotlib.

**2. READ the data values from bankloans.csv**

We import the required dataset.

**3. CREATE summary of the given data in tabular format.**

We create a basic summary of the 850 entries that have been given in the dataset. This makes it easier to understand a vast dataset. We create the columns Count, Defaulters, Sum, Mean, Median, Standard deviation, Variance, Minimum, Maximum.

**4. APPLY outlier clipping and missing imputation methods to clean data.**

Data cleaning is essential to make sure that there are no outliers present in the data. Outliers skew the accuracy of the train model and make it hard for the model to predict defaulters correctly. We clip the upper percentiles and input mean values in the missing columns to make sure that the dataset values are optimized for modelling.

**5. CREATE heatmap to display the correlation between the different variables.**

Heatmaps are used to show the relationship between two variables. The darker the box shade, the more the stronger the relationship between the two variables.

**6. DISPLAY Boxplots of each variable, comparing non-defaulters and defaulters.**

We start our data visualization by using boxplots to compare the data between defaulters and non-defaulters. By creating boxplots, we can interpret and compare the trends between defaulters and non-defaulters. We can use this to understand which parameters are needed for modelling purposes on a qualitative level.

**7. BIVARIATE DATA ANALYISIS using T-test.**

T-test is used to obtain the correlation between different variables. Using t-test we can obtain p-value which is a useful indicator for identifying the parameters impacting the model.

**8. DATA IMPORTANCE VISUALIZATION using Distplot**

Distplots combine multiple histograms and compares them. In a way it is used in the same way as Boxplots. It is utilized for qualitative comparison between defaulters and non-defaulters with respect to various parameters. It is another method for visualizing the important data parameters required for building the model.

**9. OBTAIN Variance Inflation Factor(VIF) to check for the variables necessary for modelling.**

This is done to calculate the VIF for different parameters. In combination with the p-value obtained from t-test, we can qualitatively understand the parameters necessary for modelling. A variable does not hold a considerable amount of weightage in the result modelling and analysis if their VIF is low and P-Value is high.

**10. IMPLEMENT Logistic Regression using Statistical Model.**

First our model implements logistic regression using the statistical model. Using this, we can obtain the cut-off probability, sensitivity, specificity, gini index and AUC of the model implemented. These values calculated will give us a general idea about the accuracy and predictive power of the model.

**11. CREATE Confusion Matrix for Train and Test Data and obtain Accuracy of the Model.**

We create the confusion matrix for both Train and Test Data from which a visual and quantitative value for actual and predicted true and false is obtained. From this we obtain true positive, true negatives, false positives and false negatives. Using this we can calculate the accuracy for both Train and Test Data Models.

**12. DECILE ANALYSIS**

Decile Analysis is used to make sure that the model work fine statistically. If the actual responses trend in descending order with descending deciles (sort of like a staircase), we know that the model predicts properly overall.

**13. IMPLEMENT Logistic Regression using sklearn**

To boost the accuracy of the model, we implement Logistic Regression using sklearn which is a machine learning library provided by Python.

**14. GINI Index for Train and Test Data**

We can find the Gini index for both Train and Test data to understand the predictive power of the model.

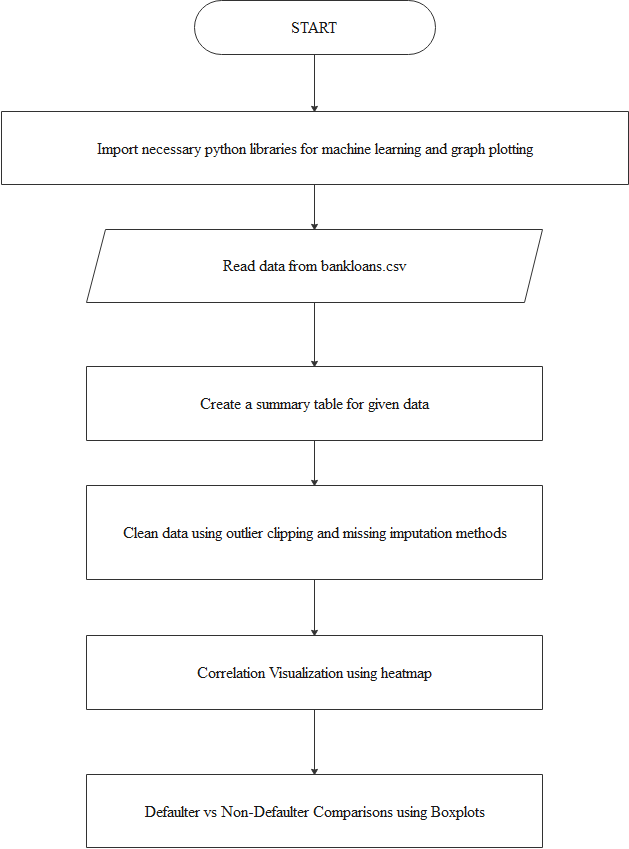
**15. CREATING AUC-ROC Curve and FINDING Cut-Off Probability**

The Receiver Operating Characteristic Curve and the associated Area Under Curve shows the effectiveness and predicting power of the model. The Cut-Off probability can also be obtained from the Curve to get optimal accuracy from the model.

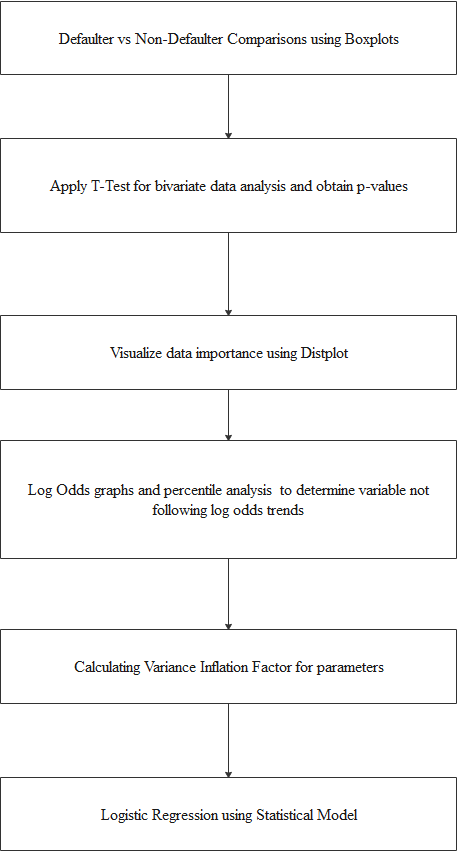
**16. ACCURACY of the model**

We find the boosted accuracy of both the Train and Test Data Models using Logistic Regression by sklearn.

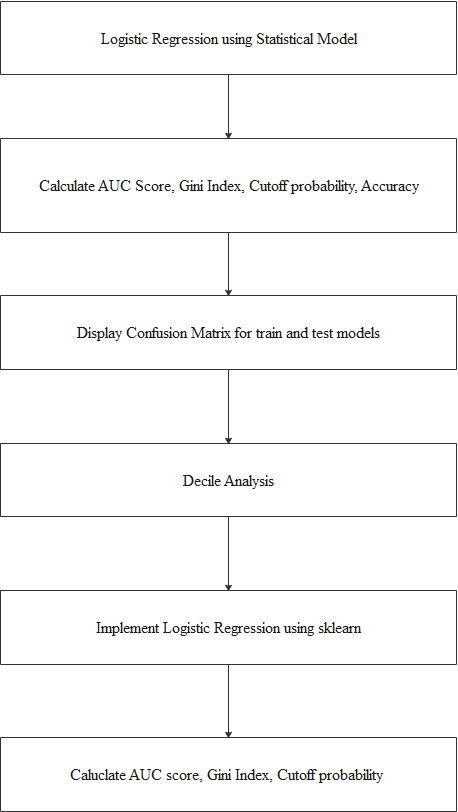
**Flowchart**



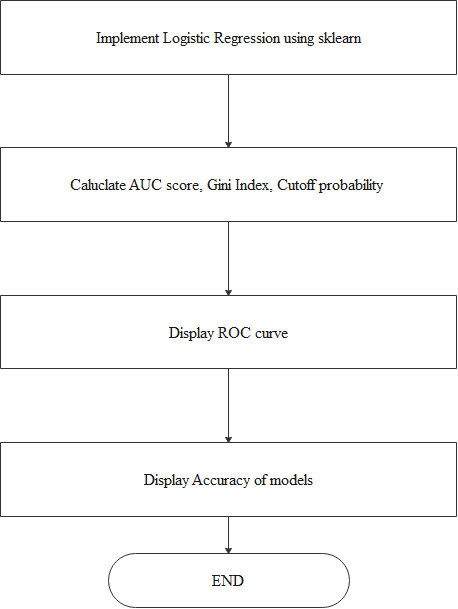
(Continued)



(continued)



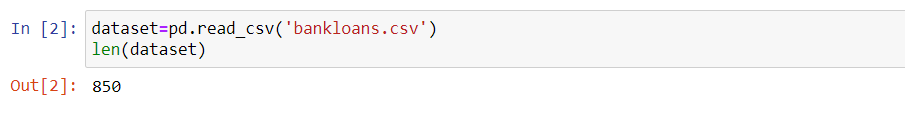
(continued)

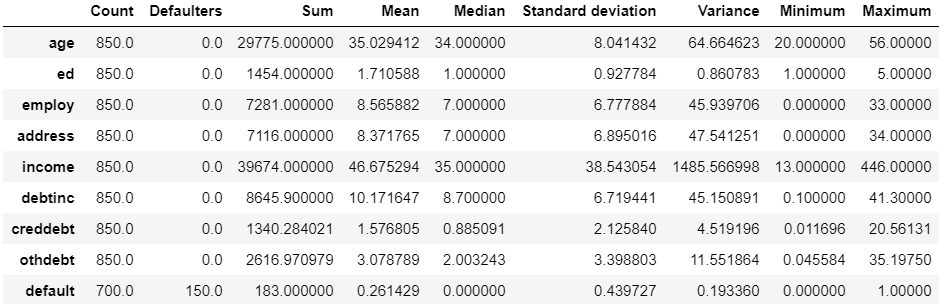


**Result Analysis**

**DATA SUMMARY and VISUALIZATIONS**

We start with loading the dataset called “bankloans.csv”. This csv files contains the details of the people who have taken a loan and were either to repay it back or were not able to replay it. It also contains several other parameters using which we will create our model. We calculate the number of entries present in the csv file. It contains 850 entries. We then find the summary of the data that is given in the dataset.





Let us look at the different parameters in the dataset:

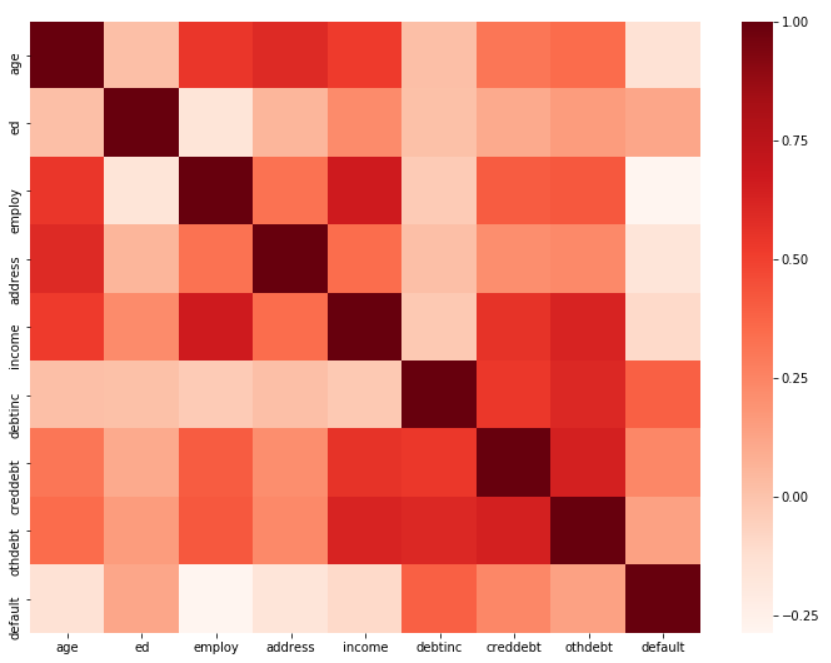
* age: This parameter represents the age of the customer.
* ed: This parameter represents the education level of the person. Education level has been quantified into levels for simplification of modelling.
* employ: This parameter shows the number of years the person has been employed continuously.
* address: This parameter shows the number of years a person has stayed at one address.
* income: This parameter represents the income of the customer.
* debtinc: This parameter represents the debt to income ratio of a particular customer.
* creddebt: This parameter represents the credit to debt ratio of a particular customer.
* othdebt: This parameter shows the other debts incurred by a customer.
* default: This parameter shows us those who have defaulted (1=defaulted, 0=not defaulted).

The columns that have been created are the following:

* Count - Number of entries in the dataset.
* Defaulters - Number of defaulters in the dataset.
* Sum – Sum of all the entries of parameters in the dataset.
* Median – Median of all the entries of parameters.
* Standard deviation – Standard deviation of the entries of parameters.
* Variance – Variance of the different entries.
* Minimum - Minimum value of the entries for different parameters.
* Maximum - Maximum value of the entries for different parameters.

**HEAT MAP**

Heat map showing the relations in our dataset:



The heat map helps us to determine the relation between each variable present in our dataset. The main variable of interest is plotted against both axis as a grid of colored squares. The darker the shade the more there is a relation between the two variables. Heat map is a way to determine correlation in a bi-variable relationship. After plotting by observing the shades across the matrix, we can observe if there is any pattern or not between them. The variables plotted on each axis can be of any type, whether they take on categorical labels or numeric values. In the latter case, the numeric value must be grouped in the form of histograms to form the grid cells where colors associated with the main variable of interest will be plotted. Cell shading corresponds to any metric, like frequency, or mean or median etc. One way of thinking of the construction of a heatmap is as a table or matrix, with color shading on top of the cells. In certain applications, it is also possible for cells to be colored based on categorical values. Categorical values are non-numerical values.

The darker the color, the stronger the relationship between those variables. Like, the relationship between variable “employ” and variable “default”, the color is very dark. This means the loan defaulter is very much dependent on the thing that if the person is employed or not. Same goes for variable employ and ed.

**DENSITY FUNCTION**

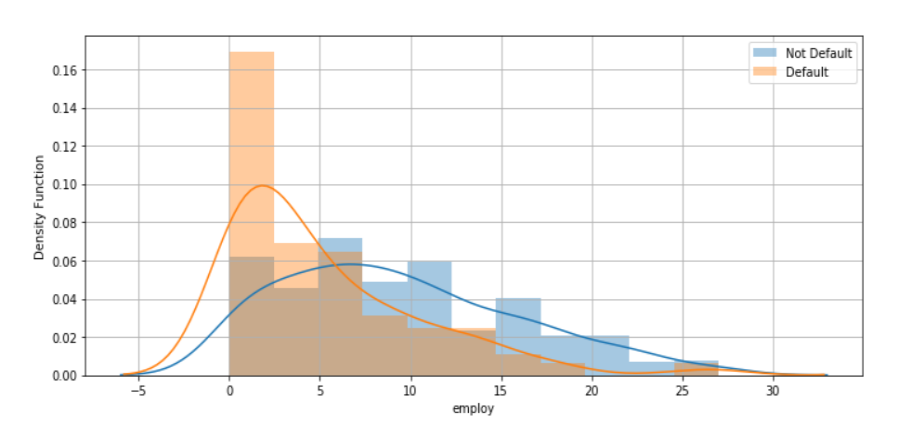
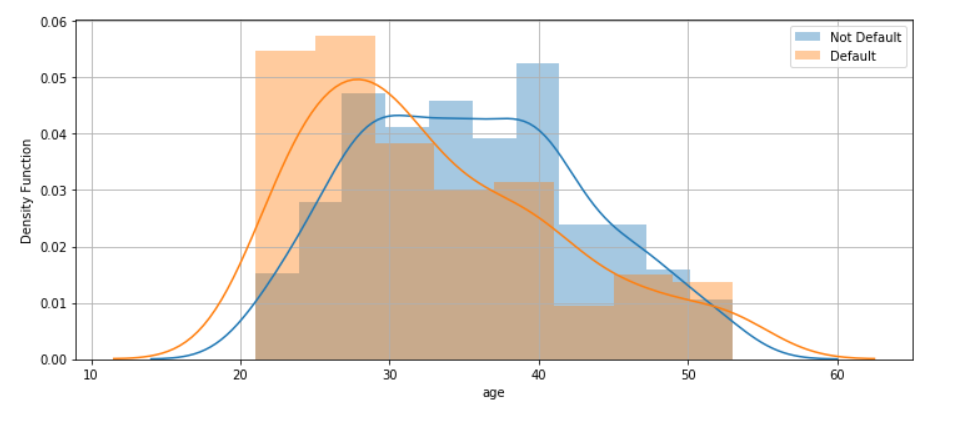
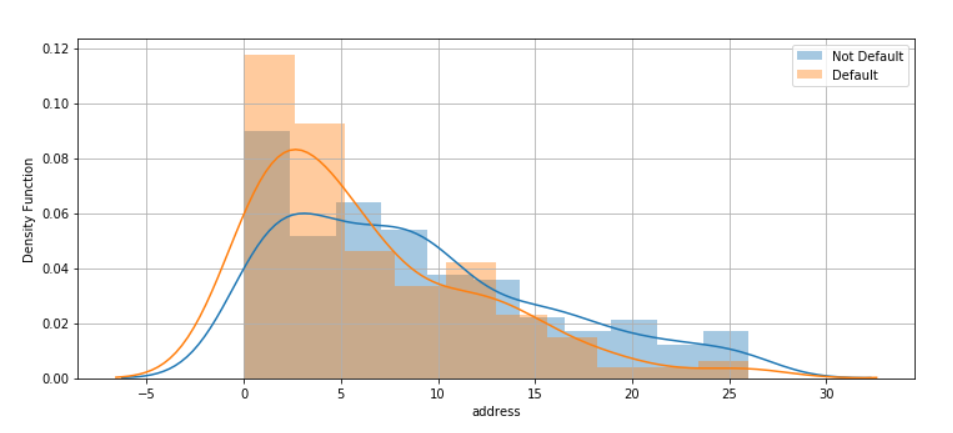
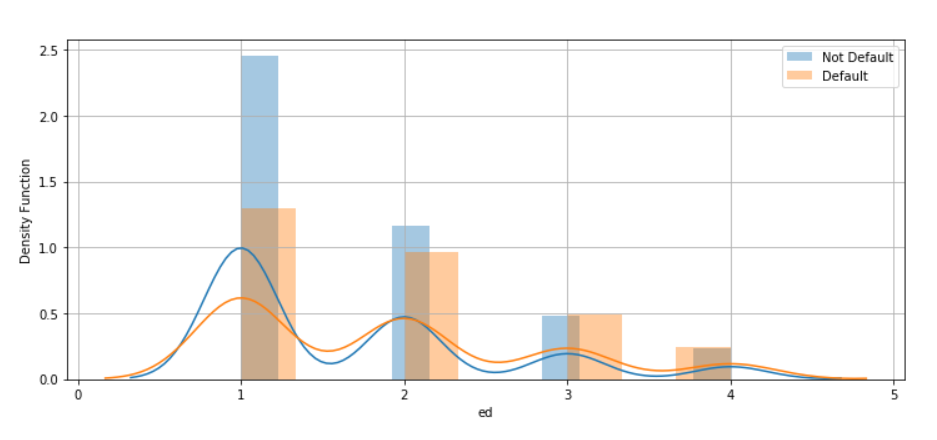
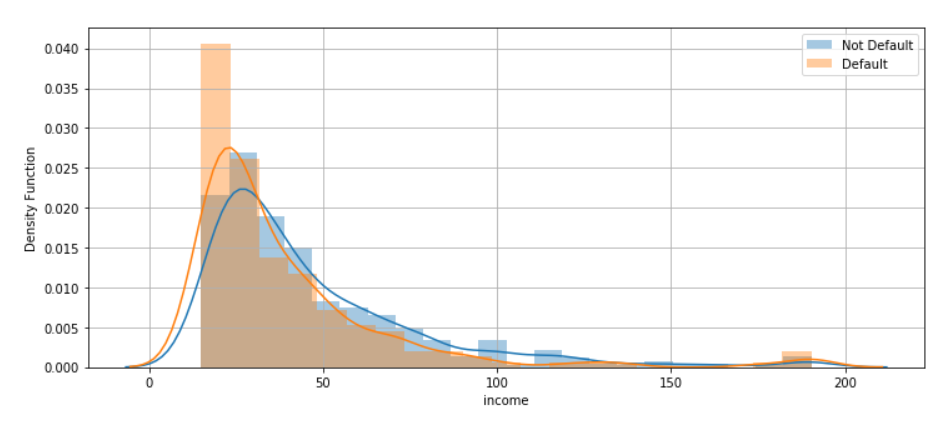
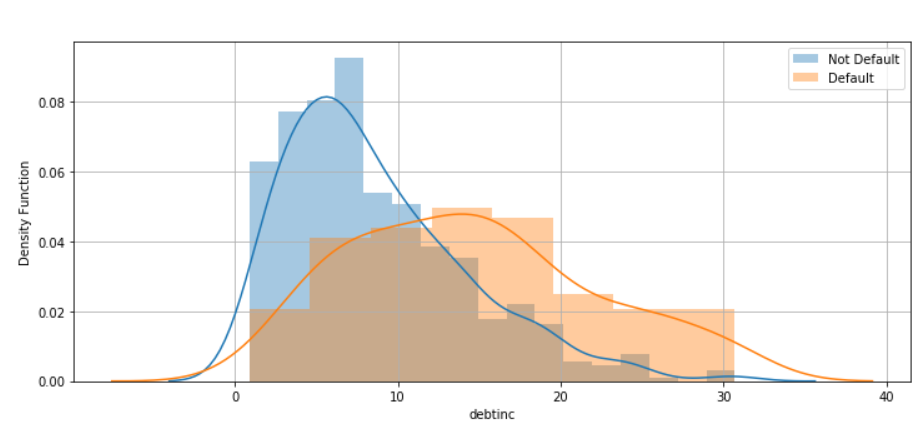
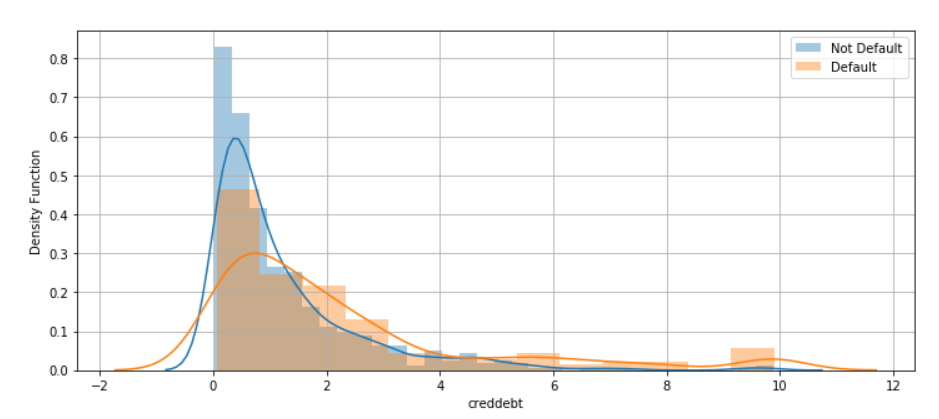
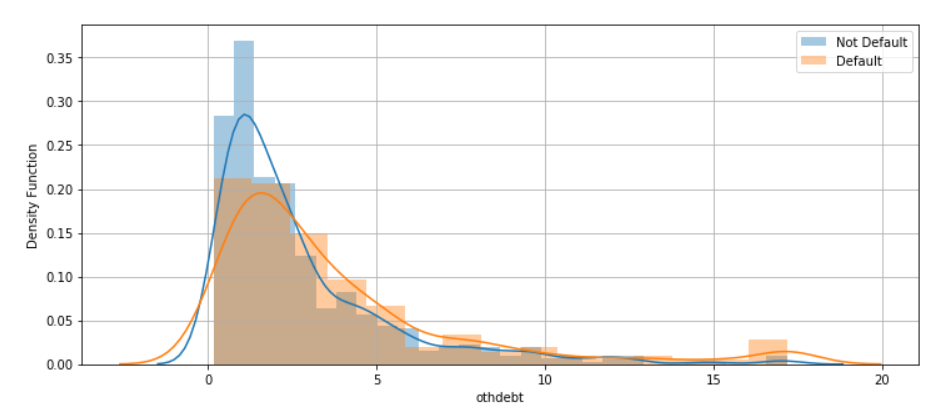
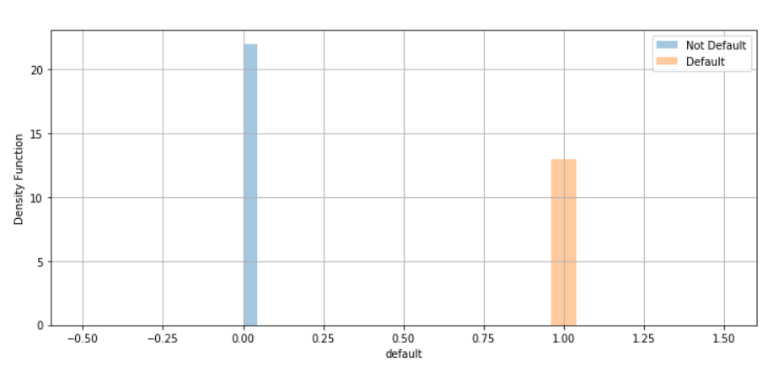
The density function analysis provides us a more detailed, precise and accurate relation of the variable fluctuation with respect to observation variance. We are plotting a separate graph of each variable with two exact plots on each of them. One is going to be ‘Default’ and the other one is ‘Non-Default’.

Default is going to convey the variation of defaulters in relation to the variable whose graph is being plotted. For example, when we are plotting a density function graph of age, the plot of Default will display the graph of people in accordance to having a chance of being a defaulter in relation to their respective age. We could have a better and continuous view of the details and variations. We could figure out the age groups with highest risk of being unable to repay the loan on time. Moreover, we can pinpoint the age numbers as well unlike grouped visualization of data. We took the data in a given range to be our observation. The extreme values in the dataset were replaced with mean values that were calculated beforehand.

We made observations of younger people, who are unemployed or having low wages of income more prone to be defaulter in loan repayment. The graph eventually declines as the age factor increases. This portrays the information of people being more reliable and less suspicious for loan repayment when their age increases. As people become settled, get employed and successfully develop a primary fixed income, the become more capable to repay their loan. These factors can clearly be taken proportional to their aging and that’s exactly what is displayed by the Non-Defaulter plot of the same graph.

Both plots are displayed under the age variable graph and thus we get a very precise display of information which is far better than grouped data visualization. Similarly, for the other graphs we can see that:

* People employed for shorter periods of time default more than those who are employed for longer periods of time. (employ graph)
* People living in one address for a log time default less than those who live less in one address. (address graph).
* Education level doesn’t impact the model much (ed graph).
* People with less income default more than people with more income. (income graph)
* People with greater debt to income ratio default more. (debtinc graph)
* People with higher credit to debt ratio default more. (creddebt graph)
* Other debts do not play much of a factor during modelling. (othdebt graph). Defaulters have slightly lower other debts.



**BOX PLOTS**

We use box plots to graphically present groups of numerical data through their quartiles. We are able to get the average value corresponding to a particular variable for both defaulters and non-defaulters separately.

1. displays the case of a defaulter.
2. displays the case of a non-defaulter.

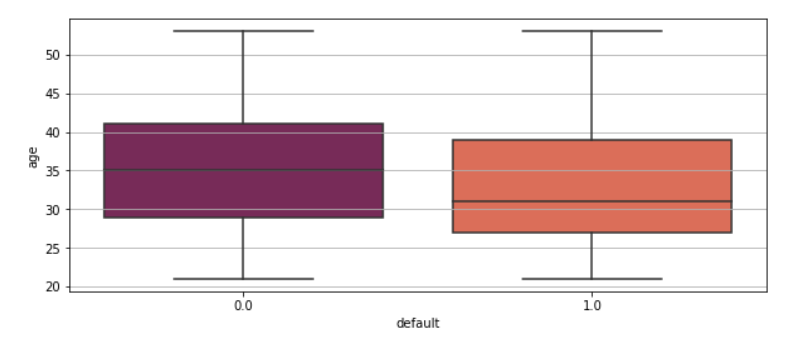
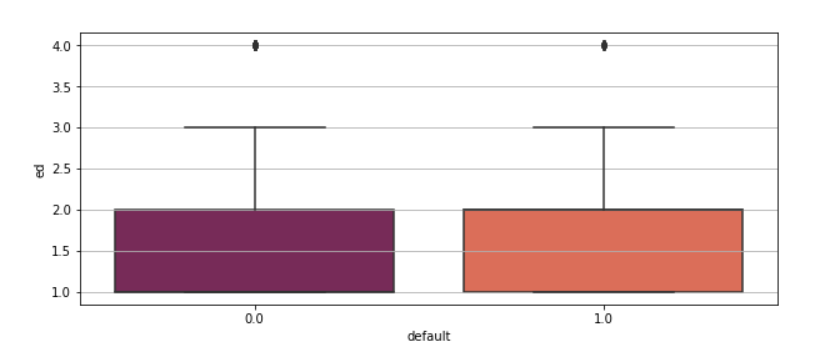
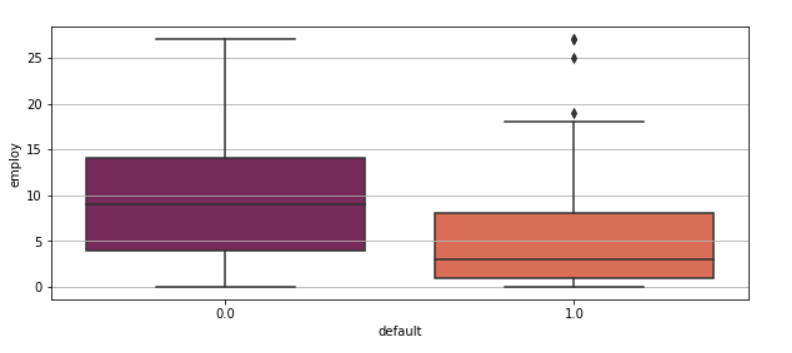
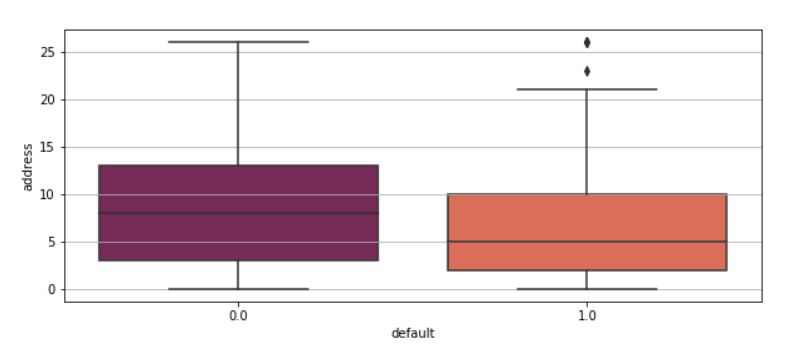
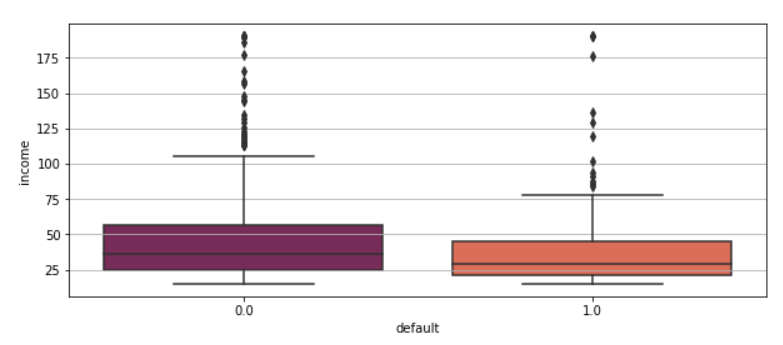
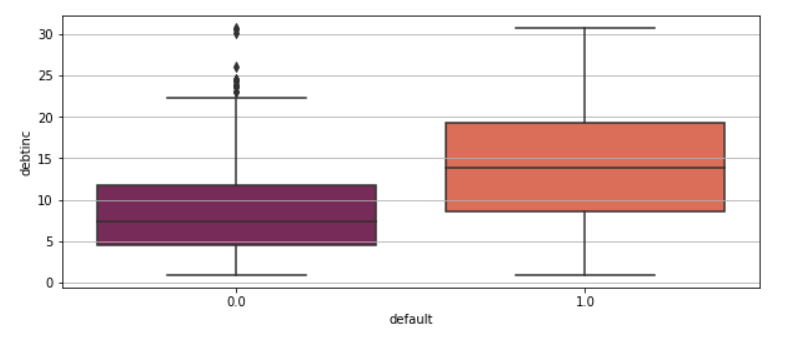
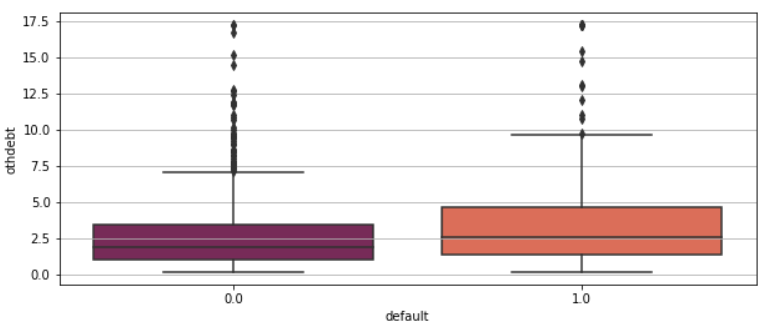
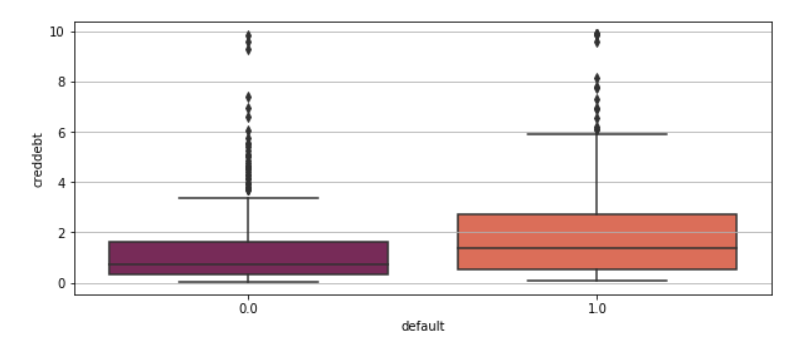
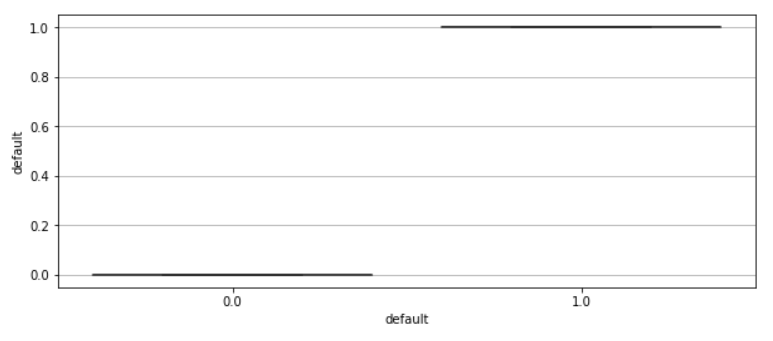
Taking the example of age, we can clearly see that the average age of a non-defaulter in this case is around thirty-six. This goes along with our hypothesis that a person becomes more financially secure with age as he develops a fixed and reliable primary source of income and thus becomes a secured borrower of loan.

On the other hand, we can observe the average of a defaulter in case of loan repayment to be around twenty. This even strengthens our hypothesis even more as young people who are unemployed or have a very low source of income tend to become a liability and an unsecured borrower of loan.

Using Box plots, we are able to calculate exact difference and prime values of fluctuation of trust in case of loan repayment for all the specific variables.

Similarly, we can see the case of debt. history. The people who have a prior loan on them are less reliable than the ones who are not having any other loan beforehand that is due. Box plots provide us with important information that also helps in analyzing the difference between the variables’ defaulter and non-defaulter ratio. We get similar comparative results as with the density plot.

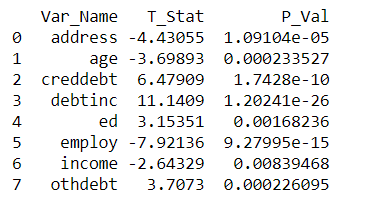
* People employed for shorter periods of time default more than those who are employed for longer periods of time. (employ graph)
* People living in one address for a log time default less than those who live less in one address. (address graph).
* Education level doesn’t impact the model much (ed graph).
* People with less income default more than people with more income. (income graph)
* People with greater debt to income ratio default more. (debtinc graph)
* People with higher credit to debt ratio default more. (creddebt graph)
* Other debts do not play much of a factor during modelling. (othdebt graph). Defaulters have slightly lower other debts.



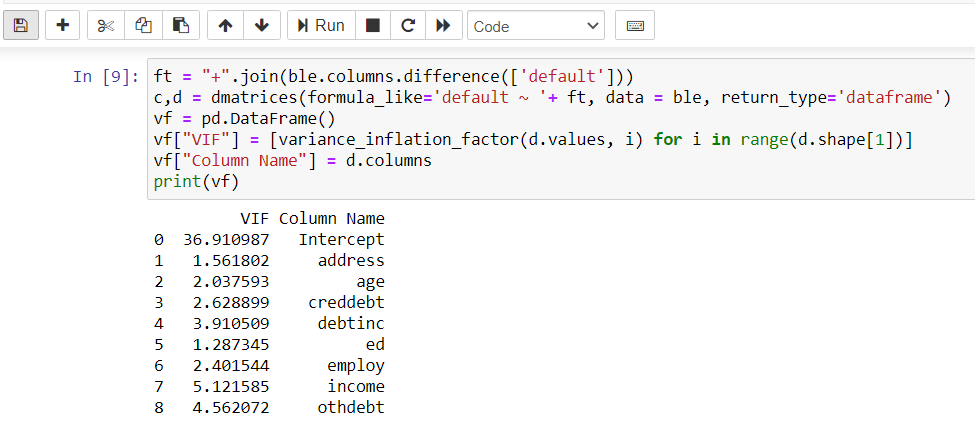
**VARIANCE INFLATION FACTOR AND P-VALUE**

The two other factors that we compute during our analysis are VIF (Variance Inflation Factor) and P-Value. P-Value acts as a factor for significance testing against null hypothesis. These values are expressed as decimals, but it is easy to read them after interpreting them into percentages. On the other hand, VIF for a regression model variable conveys the ratio of overall model variance to the variance of the model that includes only the single independent variable. These are also expressed as a decimal. They also convey an important information to us where we can see more clearly how much our final result exactly depends on a particular variable. This can be really helpful in determining whether a variable can be left out of our calculation or whether a variable is holding a heavy contribution in our final result. Though both the observations are important and go hand in hand for the calculation and a variable’s insignificance are can’t be determined on merely one of the results. From what we observe in out calculations, we conclude that perhaps age and educational qualifications do not hold such importance in our model. Or perhaps, it would be better to say that other variables might hold more importance in the result modelling and analysis. We come to conclude that a variable does not hold a considerable amount of weightage in the result modelling and analysis if their VIF is low and P-Value is high.

**P-Value**



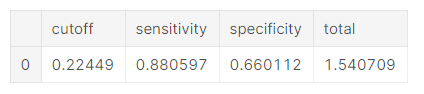
**VIF**



**LOGISTIC REGRESSION MODULE**

**I. Implementation using Statistical Model**

**1. Cutoff probability, sensitivity, specificity**



**Analysis**

Using Statistical model we have calculated the threshold or cut-off probability, sensitivity and specificity. Cut-off probability in binary classification is the probability that the prediction is true. It represents the trade off between false positives and false negatives. Logistic Regression model is optimized using an appropriate cut-off probability to achieve the best accuracy with the model. Using Statistical model, we have obtained a cut-off probability of 0.22449, which represents the fact that this probability is the most optimal threshold for the given logistic regression model and we will get the best accuracy using this cut-off probability.

Sensitivity refers to the true positive rate. True positive rate is the proportion of genuine positive results which give a positive result in a model. Having higher sensitivity is preferable as this means that we obtain truer results and more accurate results as the model can predict true positives better. The Logistic Regression model gives us a sensitivity of 0.880597 which means that it can detect true positives around 88% of the time.

Sensitivity = (Number of true positives)/(Number of true positives + Number of false negatives)

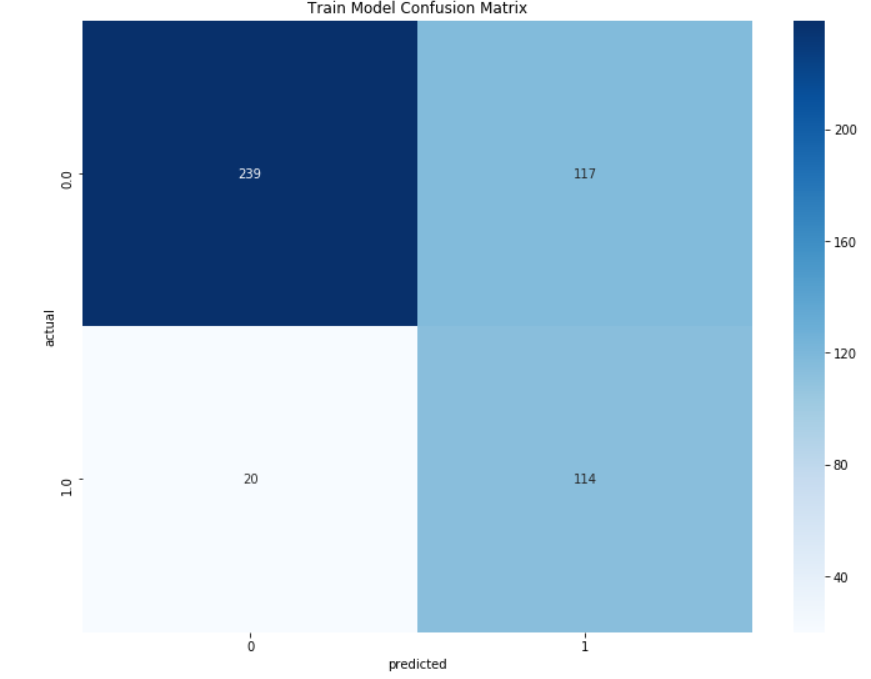
Specificity refers to the true negative rate. True negative rate is the proportion of genuinely negative results which give a negative result in the model. As with sensitivity, having a higher sensitivity value is also preferred as this implies that the model can predict the negative results more accurately, hence a higher specificity is more accurate. Logistic Regression model gives us a 0.660112 specificity value which means that true negatives are detected roughly 66% of the time.

Specificity = (Number of true negatives)/(Number of true negatives + number of false positives)

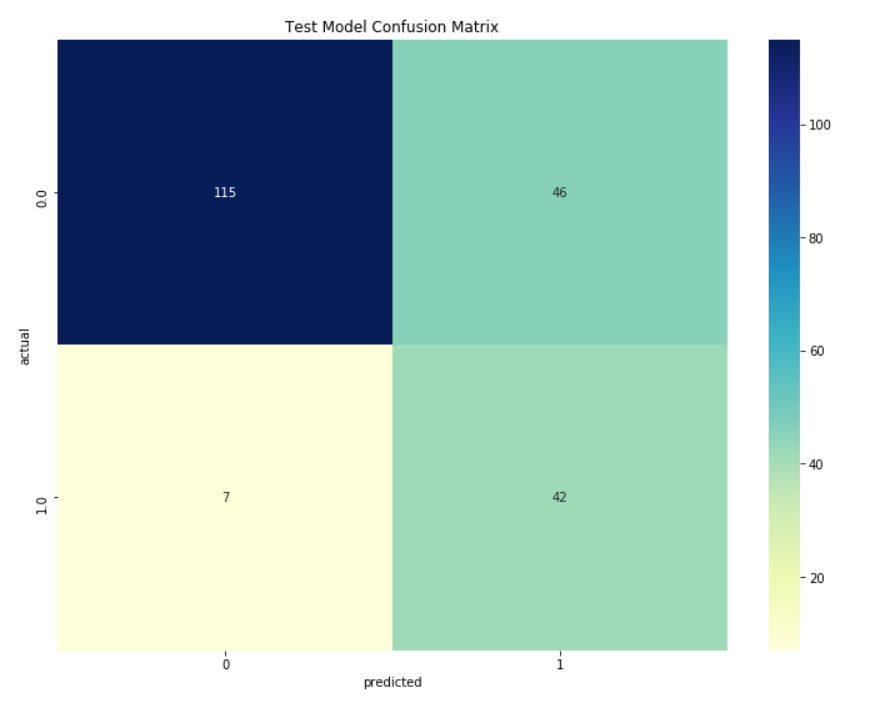
Total refers to the addition of sensitivity + specificity. Higher this value implies the more accurate the results. In the model, we obtained max total value of 1.540709 and that implies we take the corresponding cut-off, sensitivity and specificity values to obtain the best results.

**2. Confusion Matrix**

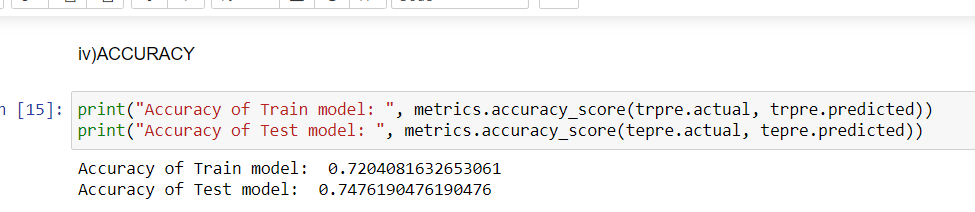
**i. For Train Data**



**ii. For Test Data**



**3. Accuracy of Logistic Regression using Statistical model**



**Analysis**

A confusion matrix is a matrix that is used to describe the performance of a classifier which in this case is Logistic Regression. Confusion matrix is a table with 4 different combinations of predicted and actual values. Using these values we can obtain specificity, accuracy, and the receiver operating characteristic curve to calculate the are under the curve. Confusion matrix is divided into 4 parts. The vertical part gives us either actual 0 or actual 1. The horizontal part gives us the predicted 0 or predicted 1. This implies that first square or section which is (0,0) gives us the true negative results. The second square which is (1,0) gives the false positive results. the third square which is (0,1) gives false negative results. The fourth square which is (1,1) gives the true positive results. True negative implies that model predicted false and it is actually false. False negative implies that model predicted false but it is actually true. False positive implies that model predicted true but it is actually false. True positive implies that model predicted true and it is actually true.

For the Train data, we can infer from the Train data confusion matrix the following details.

* True Negative = 239
* False Positive = 117
* False Negative = 20
* True Positive = 114

For the Test data, we can infer from the Test data confusion matrix the following details.

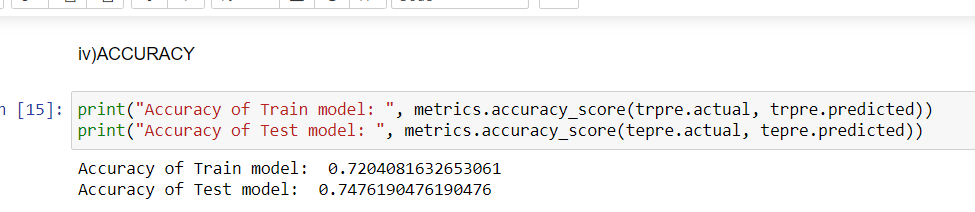
* True Negative = 115
* False Positive = 46
* False Negative = 7
* True Positive = 42

We have obtained the confusion matrices for both the Train data as well as for the Test data. As we have the confusion matrices for both of them, it is now possible for us to calculate the accuracy of both the train and test data.

Accuracy is given as : (True Positive + True Negative)/ (Total)

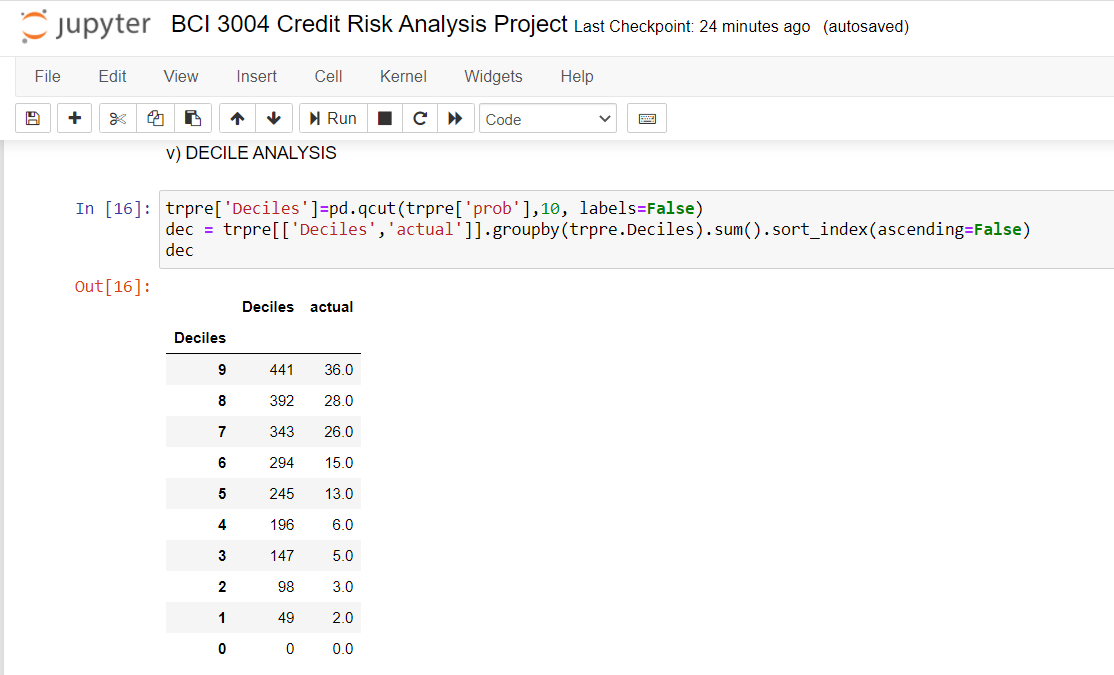
* Accuracy of Train model is : (114+239)/(239+117+20+114) = 0.7204081633
* Accuracy of Test model is: (115+42)/(115+46+7+42) = 0.7476190476

The results we have calculated here are substantiated by the results we have obtained in the accuracy of our models.



We have thus obtained roughly 72% accuracy for the train data and 74.76% accuracy for the test data. These are considered to be good values for accuracy when implementing via Statistical model. However, we can further boost the accuracy of our model by implementing Logistic Regression using sklearn which is a python machine learning library.

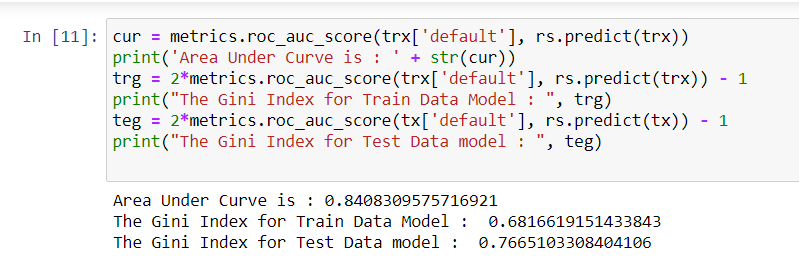
**II. Decile Analysis**



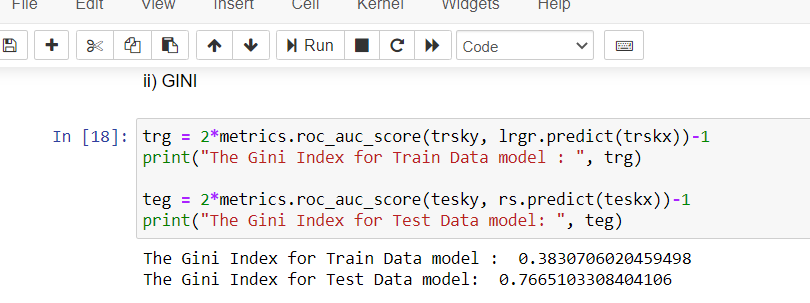
Decile analysis is a tool which is used to validate a classifier which in this case is logistic regression. A decile analysis is used to test the model’s ability to predict the required result. Decile analysis is done by creating a validation sample which is then scored according to the model that is being tested. The data is sorted in descending order and it is deivided into 10 parts known as deciles based on predicted outcomes. The top decile contains the 10% of the dataset values that are most likely to default while the lowest decile contains 10% of the the data values that will not default. The actual column represents the number of defaulters in a given decile. When using decile analysis, the ideal situation is to obtain a staircase effect. This means that the model is segregating the entries correctly in terms of most likely to default to least likely to default. As we can see in our decile analysis table, We can see a descending order trend being followed in both the decile value as well as in actual defaulters value. This implies that our model is working properly statistically and the segregation into deciles has been accurately.

**III. Gini Index for Logistic Regression**

i. Statistical model

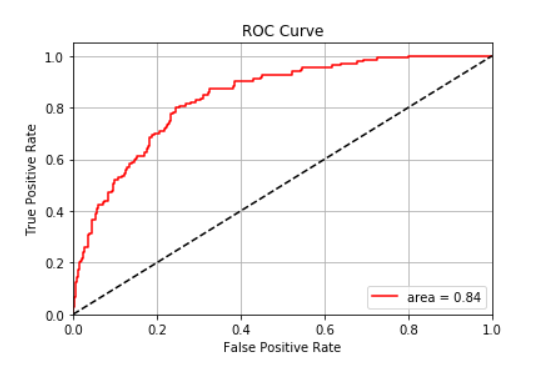


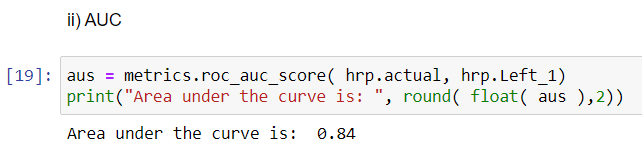
ii. sklearn

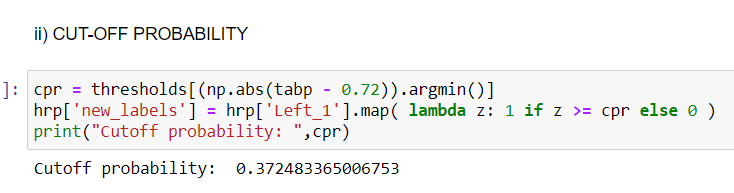


Gini Index is a very popular evaluation metric for rating credit models. It indicates the models discriminatory or inequality power which indicates the effectiveness of the model in differentiating between those who have greater chance to default and those who wont default in the future. It is a means to evaluate prediction power of a particular model. It is a measure of the ordinal relationship between two variables. The Gini Index can take on a value between -1 and 1. Here -1 represents a perfect negative ordinal relationship and 1 represents a perfect positive ordinal relationship. 0 means that there is no relationship. In practice a Gini Index of 0.4 is considered to be optimal for creating credit scoring models. It is good practice in credit models to get a positive relationship between variables which we have obtained in both the Statistical and sklearn methods of implementing Logistic Regression. The Gini Index is mostly used in imbalanced datasets where it is not possible to forecast or predict outcomes based on probability alone. It is used in risk models since the probability of default is generally low. Using sklearn we obtain train data gini as 0.38 which implies that it has 38% more predictive power than a native or random model. As the test data sample size is smaller, we obtain a greater gini of around 76% which implies greater predictive power of test model.

**IV. AUC, ROC and Cut-off probability**







**Analysis**

ROC stands for receiver operating characteristic curve. It is used in logistic regression to get the optimal cut-off value for predicting whether an entry is false(0) or true(1). It is often used for evaluation of logistic regression models. It is a good way to view how a model differentiates between true positives and false positives. The ROC curve plots sensitivity of the model as True Positive Rate(TPR) vs False Positive Rate(FPR). In general, the higher the deviation of the ROC curve from the 45 degrees line, the more the predictive power of that particular model. The 45-degree line implies 0 predictive power. An ROC curve plots TPR vs FPR at various thresholds. Lowering the threshold or cut-off probability increases the classification of entries as true hence increasing the number of true positives and false positives.

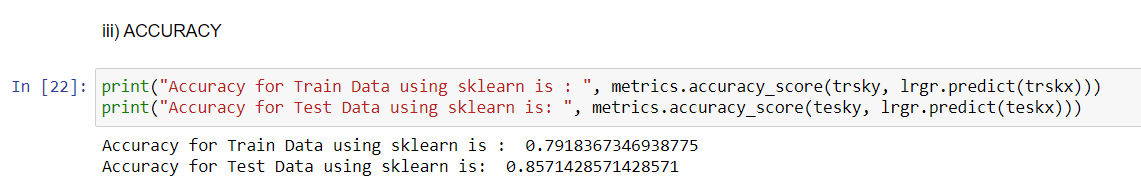
The Area Under Curve (AUC) can be calculated from the ROC curve. AUC measures the area between the x-axis and the curve. AUC will always lie between 0.5 and 1 as the ROC curve will never go below the 45-degree line. The greater the AUC, the higher the predictive power of the model. AUC of 1 implies perfect predictive power. AUC provides an average measure of the performance of the model across different cut-offs.

Cut-off Probability can be obtained from the ROC curve. We know that the ROC curve plots TPR vs FPR where TPR is sensitivity. We want that cut-off probability which has the highest total sensitivity and specificity. We can determine the optimal cut-off probability from ROC curve.

In our model we have obtained:

* An ROC curve which has considerable deviation from the standard 45-degree line. Thus implies that the model has considerable predictive power.
* AUC = 0.84. The value of AUC is quite high and this shows that the model performs well and has good predictive power. AUC is cut-off invariant and it gives a general overview of effectiveness of model.
* Cut-off probability = 0.366353. This is the cut-off probability that gives us the maximum sensitivity and specificity for the model. This threshold value will give the most optimal accuracy score for the model.

**V. Accuracy of Logistic Regression using sklearn**



We can see that we have obtained a greater accuracy score for both Train and Test data as compared to the statistical model. Sklearn Logistic Regression allows us to obtain the accuracy directly without the requirement of confusion matrix. This method is more effective and has greater predictive power as compared to the statistical model.

* Accuracy of Train Data Model using sklearn = 79.1%
* Accuracy of Test Data Model using sklearn = 85.7%

**CONCLUSION**

In conclusion, we see that credit risk analysis is an extremely important and relevant real-world topic which could affect multiple businesses. We see the importance of understanding the data which is presented to the model for training purposes. It is important to get an overview of given data and to optimize it to obtain better accuracy. To achieve this, we have implemented outlier clipping and mean, median imputation techniques to maintain the consistency techniques. It is also important to explore the data that is given. We have achieved this by creating a heatmap which shows the relation between two different parameters and extends it to all parameters in a matrix of different shades where the darker shades represent stronger relationship between two parameters. We have also dived into qualitative data analysis and visualization by means of boxplots and distplots both of which give us an overview of the trends followed by defaulters and non-defaulters with respect to various parameters. Quantitative correlation and importance in modelling has been achieved by means of statistical tests such as T-test and by calculation of VIF to determine the multicollinearity.

The model is first created by using Statistical methods by use of Logistic Regression Classifier and we obtain the Cut-off probability, sensitivity, specificity of the model. We also get the gini index and AUC values and also create confusion matrices for train and test data from which we get a respectable accuracy of 72% for Train data and 74.7% for Test data models, respectively.

In order to boost the accuracy of our model, we implement Logistic Regression by means of sklearn and obtain the gini index, AUC and create an ROC curve and this time we get a boosted accuracy of 79.1% for Train data and 85.7% for Test data models respectively. From this we can see that we have created a model which has 85.7% accuracy in predicting defaulters. Thus we can see the huge impact that using sklearn makes to the overall accuracy of the model.

**REFERENCES**

[1] M. Becha, O. Dridi, O. Riabi, and Y. Benmessaoud, “Use of Machine Learning Techniques in Financial Forecasting,” pp. 1–6, 2020, doi: 10.1109/octa49274.2020.9151854.

[2] M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M. J. Segovia-Vargas, “Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending,” *IEEE Access*, vol. 8, pp. 64873–64890, 2020, doi: 10.1109/ACCESS.2020.2984412.

[3] G. Caruso, S. A. Gattone, F. Fortuna, and T. Di Battista, “Cluster Analysis for mixed data: An application to credit risk evaluation,” *Socioecon. Plann. Sci.*, no. November 2019, p. 100850, 2020, doi: 10.1016/j.seps.2020.100850.

[4] K. Alshouiliy, A. Alghamdi, and D. P. Agrawal, “AzureML based analysis and prediction loan borrowers creditworthy,” *Proc. - 3rd Int. Conf. Inf. Comput. Technol. ICICT 2020*, pp. 302–306, 2020, doi: 10.1109/ICICT50521.2020.00053.

[5] W. Zhang, C. Wang, Y. Zhang, and J. Wang, “Credit risk evaluation model with textual features from loan descriptions for P2P lending,” *Electron. Commer. Res. Appl.*, vol. 42, p. 100989, 2020, doi: 10.1016/j.elerap.2020.100989.

[6] X. Yu, Q. Yang, R. Wang, R. Fang, and M. Deng, “Data cleaning for personal credit scoring by utilizing social media data: An empirical study,” *IEEE Intell. Syst.*, vol. 35, no. 2, pp. 7–15, 2020, doi: 10.1109/MIS.2020.2972214.

[7] V. Hassija, G. Bansal, V. Chamola, N. Kumar, and M. Guizani, “Secure Lending: Blockchain and Prospect Theory-Based Decentralized Credit Scoring Model,” *IEEE Trans. Netw. Sci. Eng.*, vol. 4697, no. c, pp. 1–10, 2020, doi: 10.1109/TNSE.2020.2982488.

[8] A STUDY ON CREDIT RISK MANAGEMENT IN SCHEDULED BANKS M.Rajeswari, Associate Professor, Department of Management Studies, Arupadai Veedu Institute of Technology, Chennai-Tamil Nadu, India

[9] CREDIT RISK MANAGEMENT IN COMMERCIAL BANKS

Konovalova N., Kristovska I., Kudinska M.\*

[10] Impact of risk management strategies onthe credit risk faced by commercial banks of Balochistan Zia Ur Rehman, Noor Muhammad\*, Bilal Sarwar and Muhammad Asif Raz

[11] The Relationship Between Credit Risk Management Practices And Profitability In Malaysian Commercial Bank`S

Ali Taha Oleiwi, Maimuna Ali, Sarmad Hamza Jassim, Mohammed Hayder Nadhim, Ganama Moustapha Gueme, Nazarudin Bujang

[12] What’s wrong with modern credit risk management?

PhD, Ihor Voloshyn

[13] Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer

1stWaseem Ahmad Chishti 2nd Shahid Mahmood Awad

[14] P. Giudici and L. Parisi, “CoRisk: Credit risk contagion with correlation network models,” Risks, vol. 6, no. 3, pp. 1–19, 2018, doi: 10.3390/risks6030095.

[15] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, “Credit risk prediction in an imbalanced social lending environment FinanceIT Research Group , University of New South Wales , Centre for Artificial Intelligence , University of Technology Sydney ,” Int. J. Comput. Intell. Syst., vol. 11, no. 1, p. 925, 2018, doi: 10.2991/ijcis.11.1.70.

[16] P. Monnin, “Integrating Climate Risks into Credit Risk Assessment - Current Methodologies and the Case of Central Banks Corporate Bond Purchases,” SSRN Electron. J., no. December, 2019, doi: 10.2139/ssrn.3350918.

[17] Y. Cui, S. Geobey, O. Weber, and H. Lin, “The impact of green lending on credit risk in China,” Sustain., vol. 10, no. 6, pp. 1–16, 2018, doi: 10.3390/su10062008.

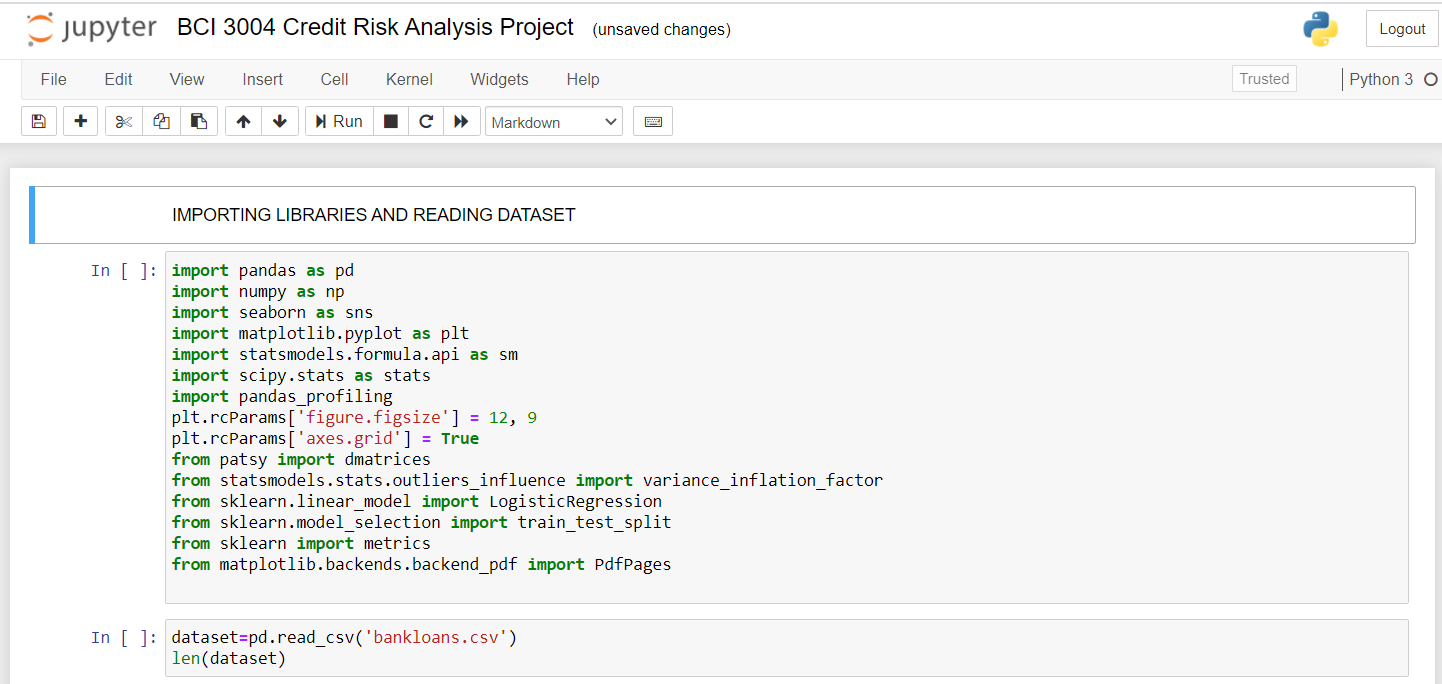
[18] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” Risks, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

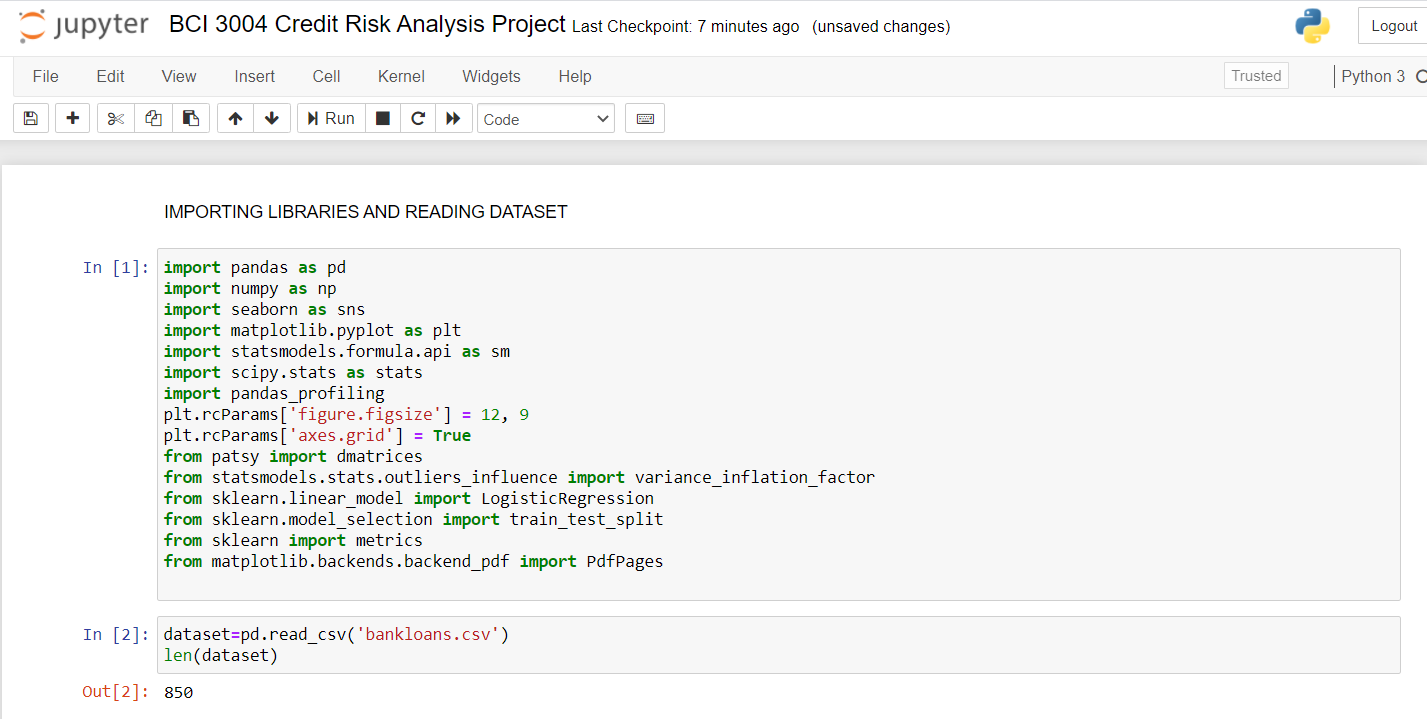
[19] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, “An Interpretable Model with Globally Consistent Explanations for Credit Risk,” pp. 1–10, 2018, [Online]. Available: http://arxiv.org/abs/1811.12615.

**APPENDIX**

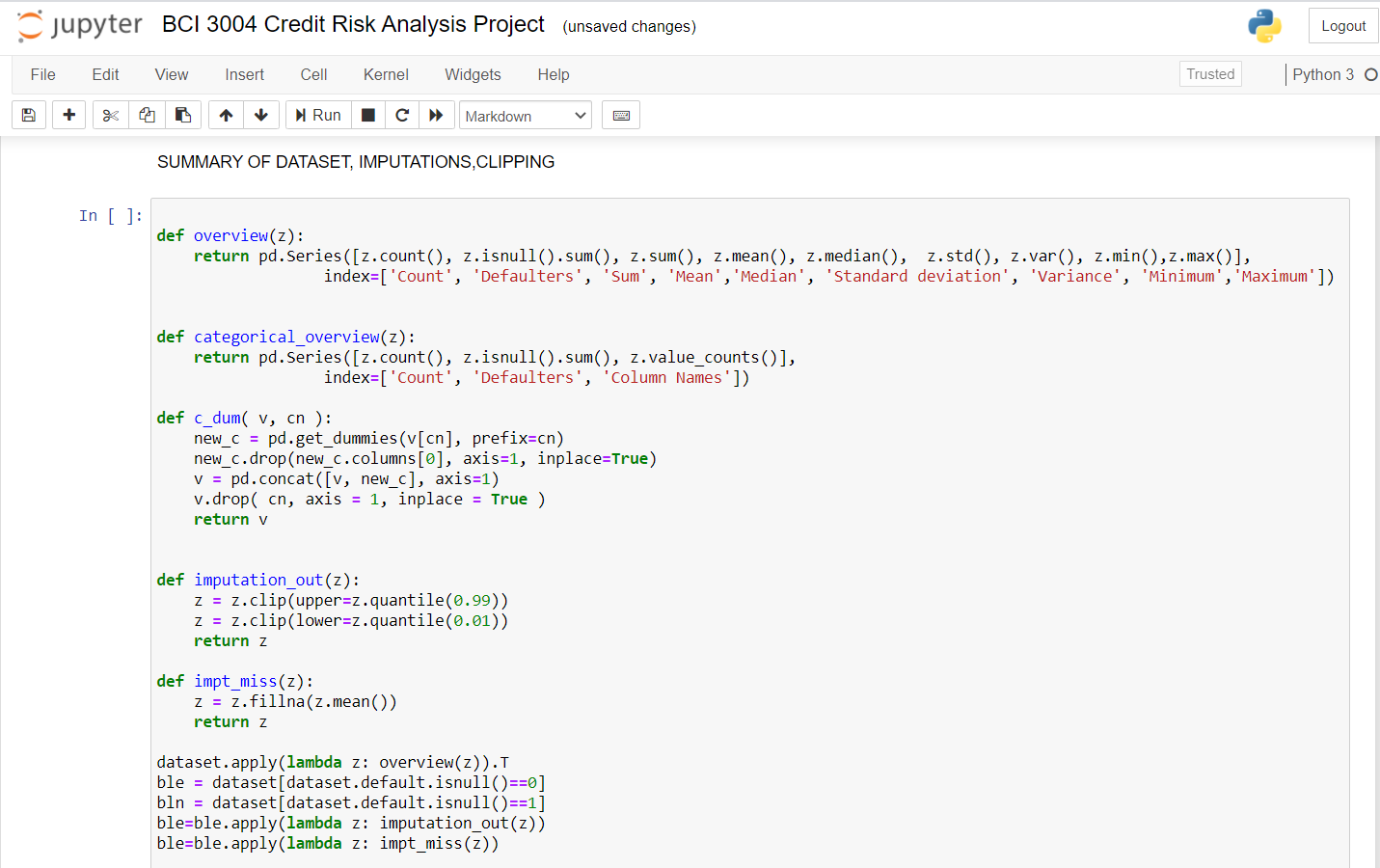
**CODE WITH MODULES ( with screenshots of output ):**

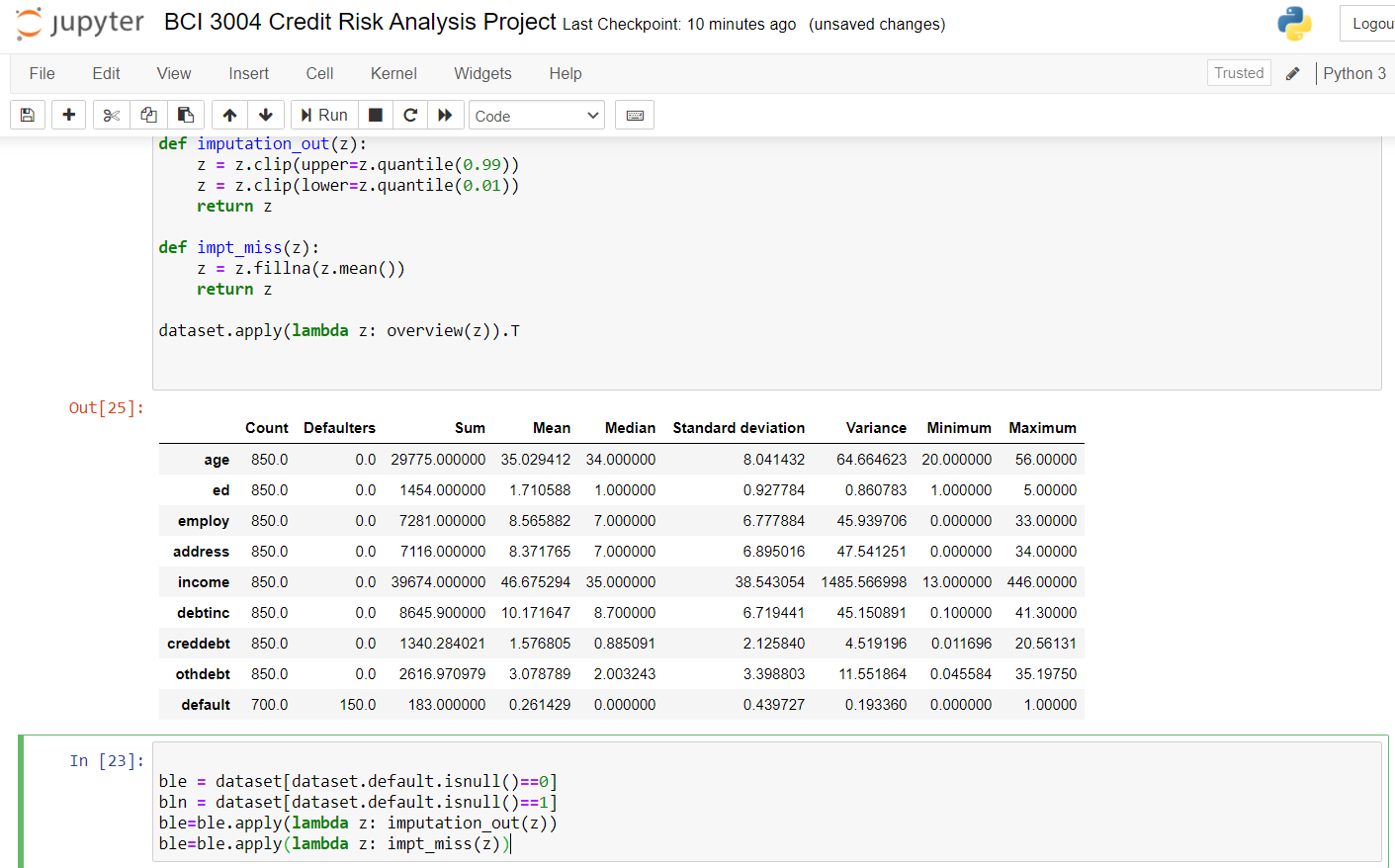
**1. Importing Libraires , Reading the dataset and finding the number of entries**

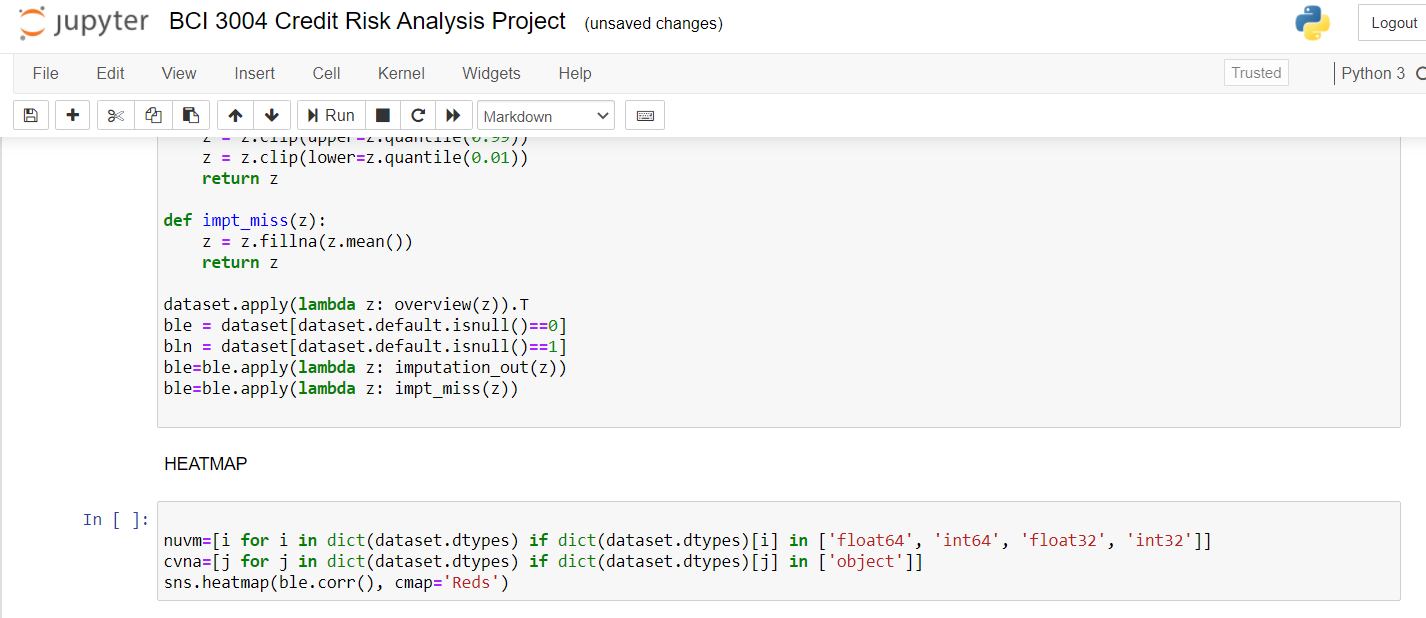


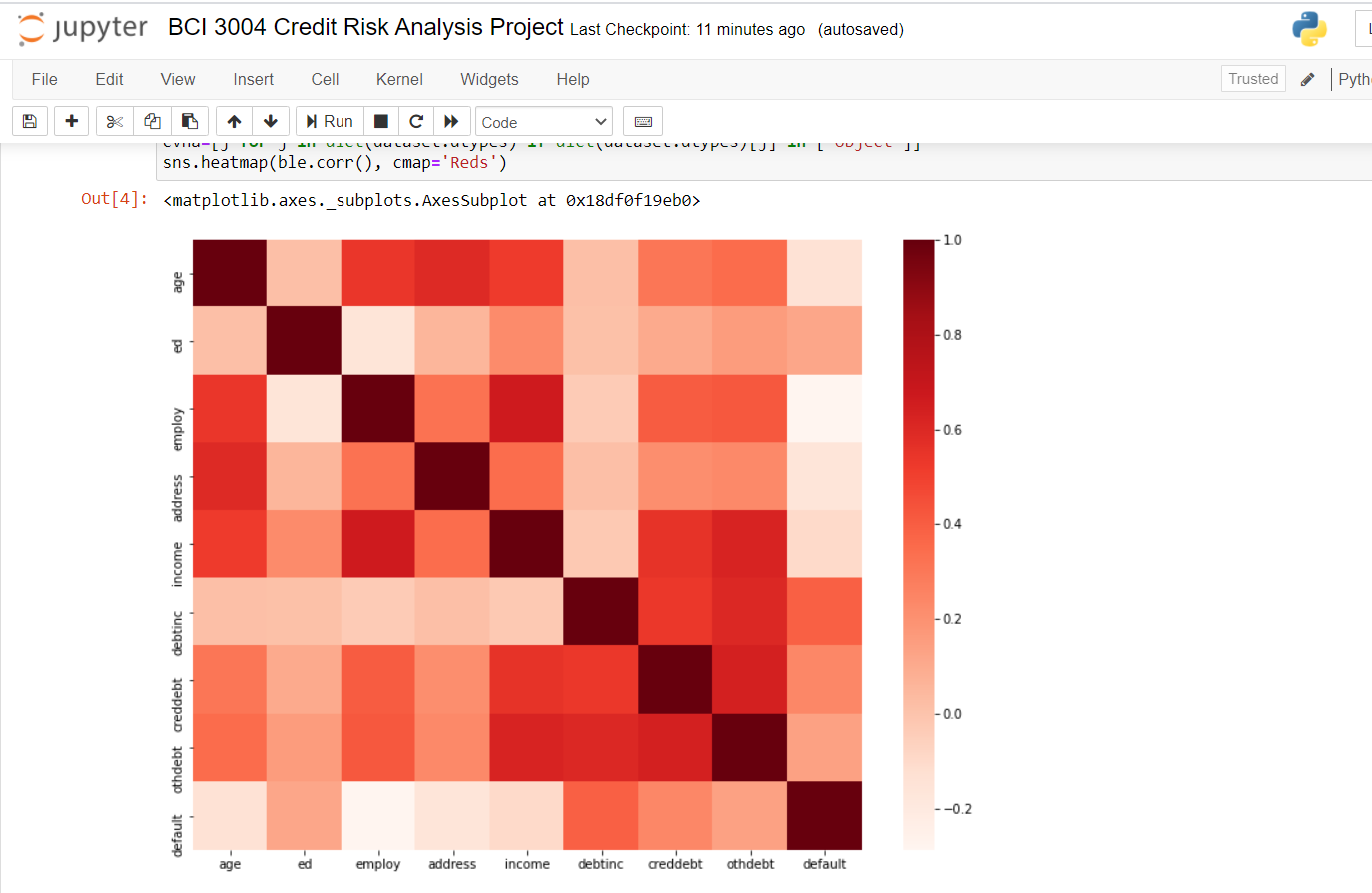


**2. Summary of dataset entries, Outlier Clipping and Heatmap**

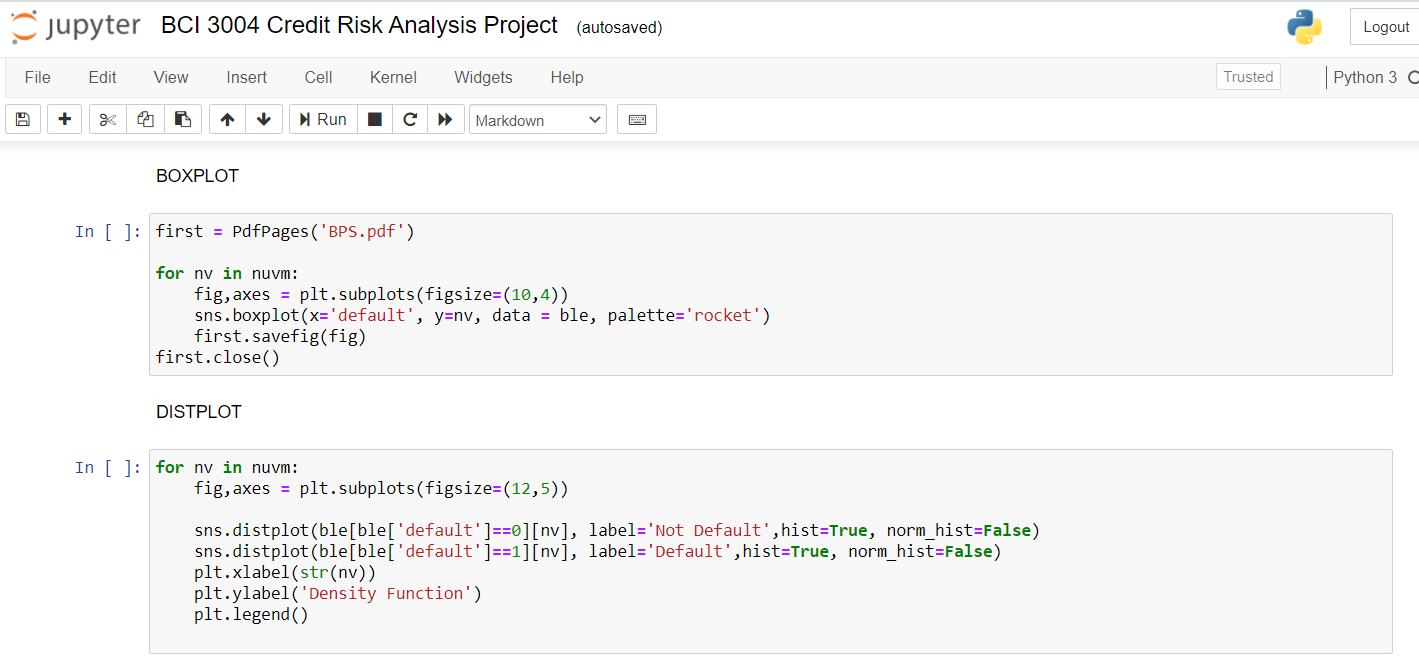


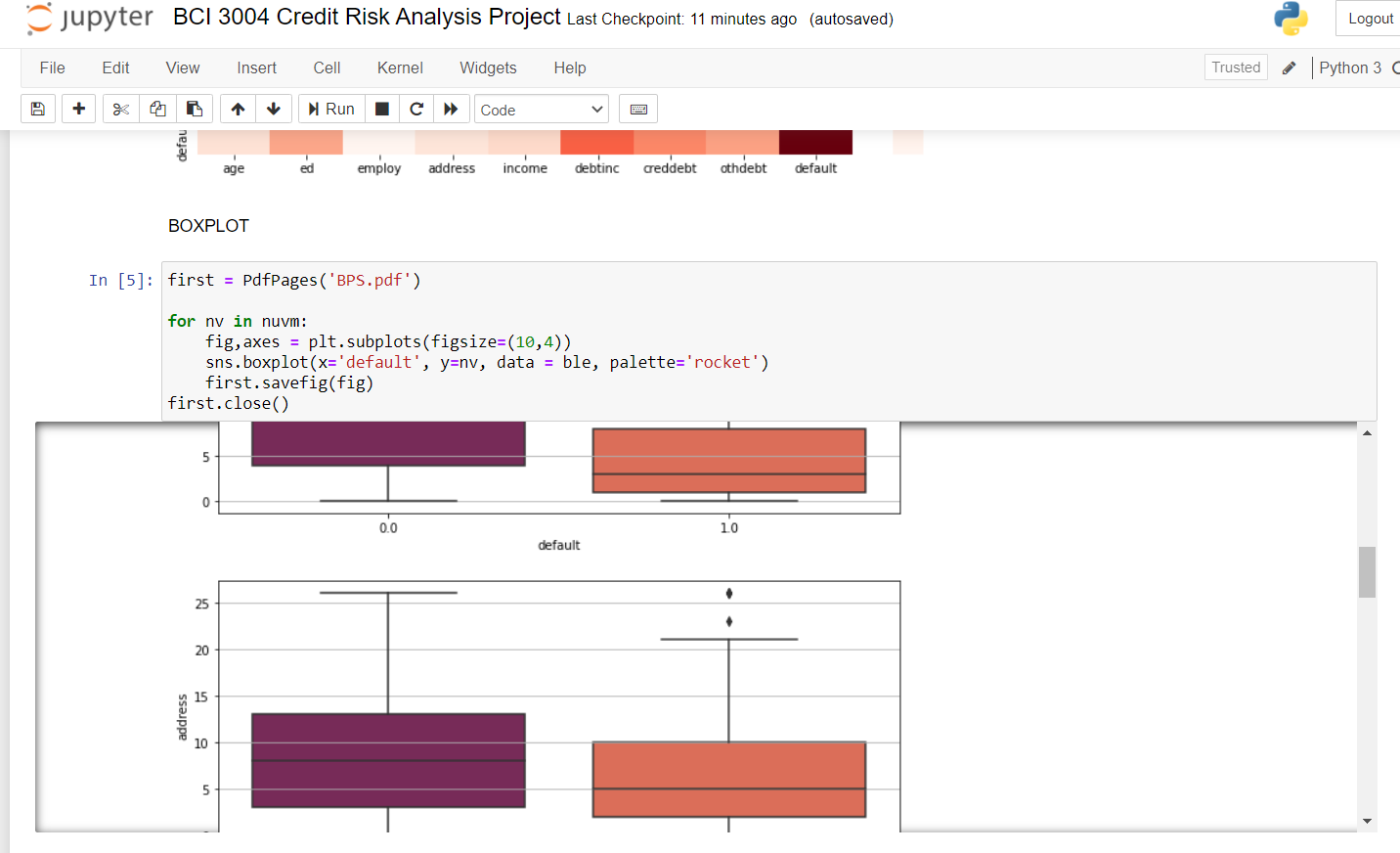


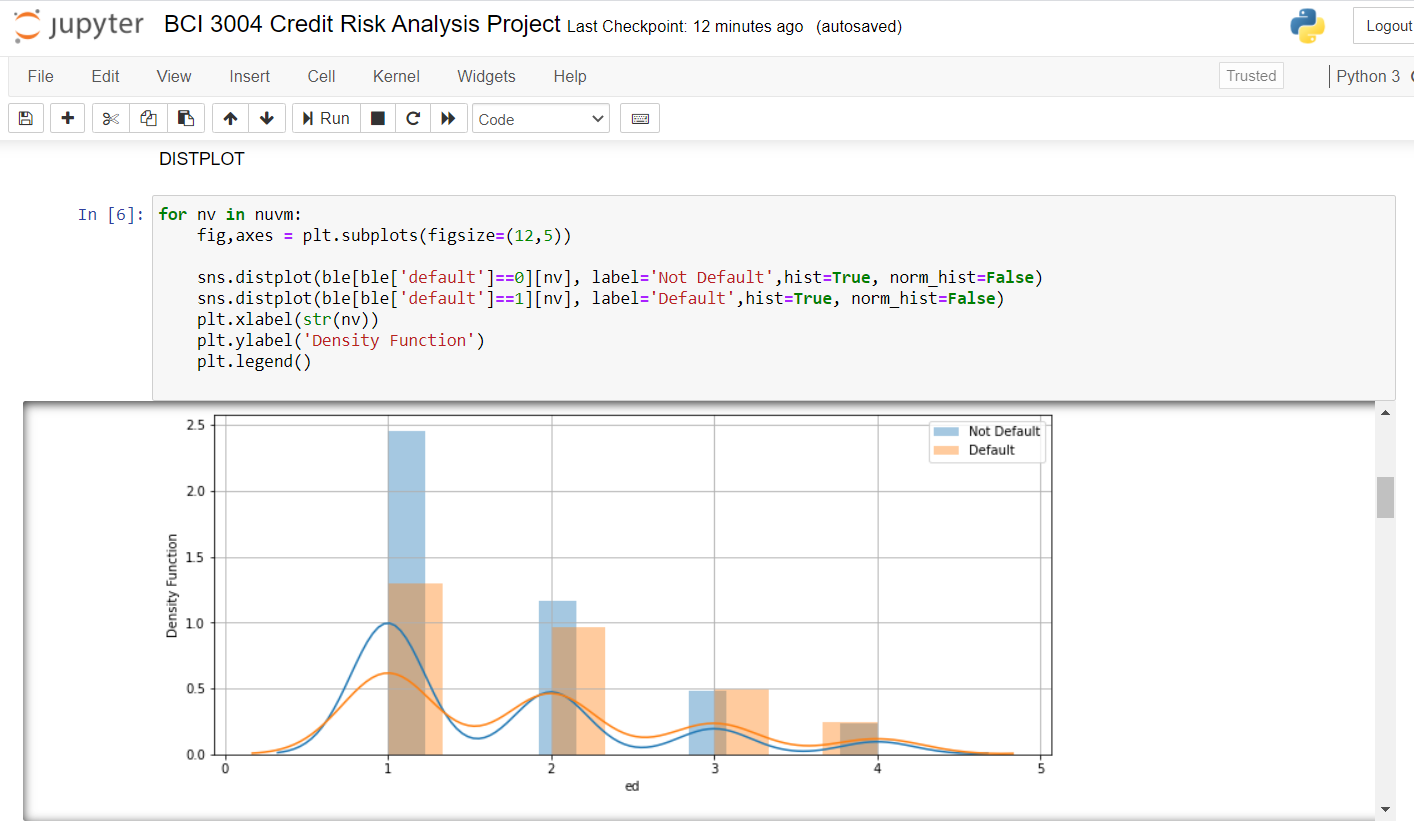




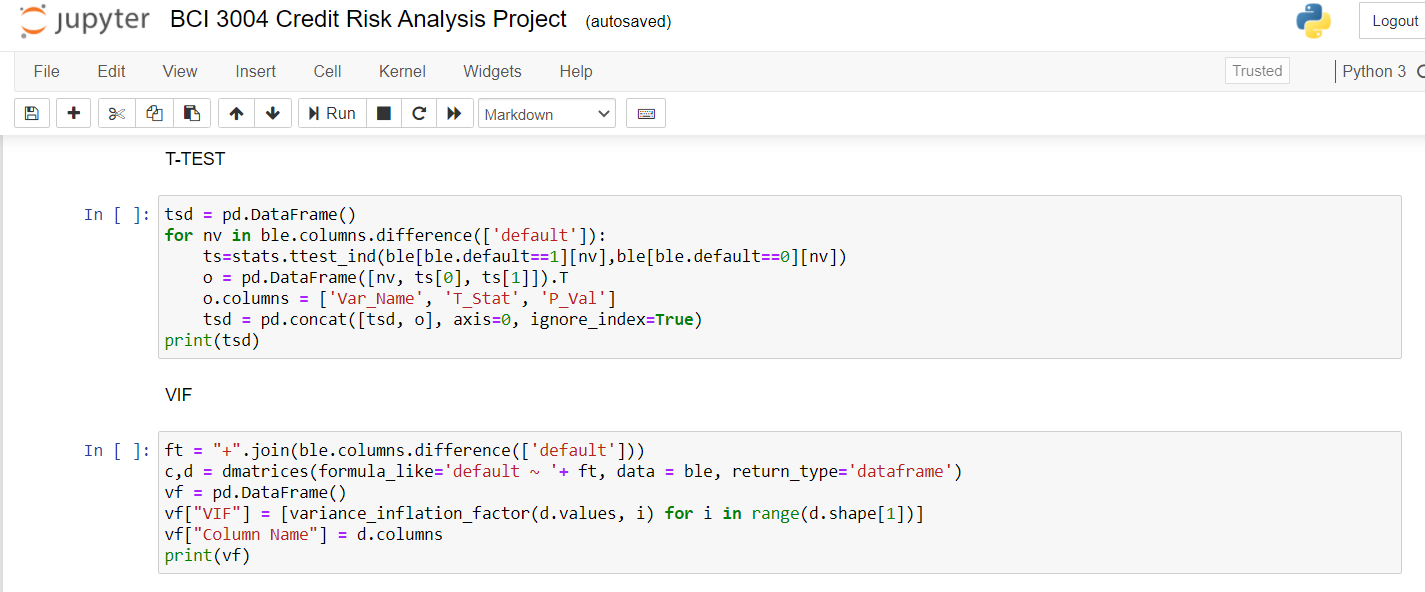
**3. i) Data visualization using BoxPlot, Distplot**

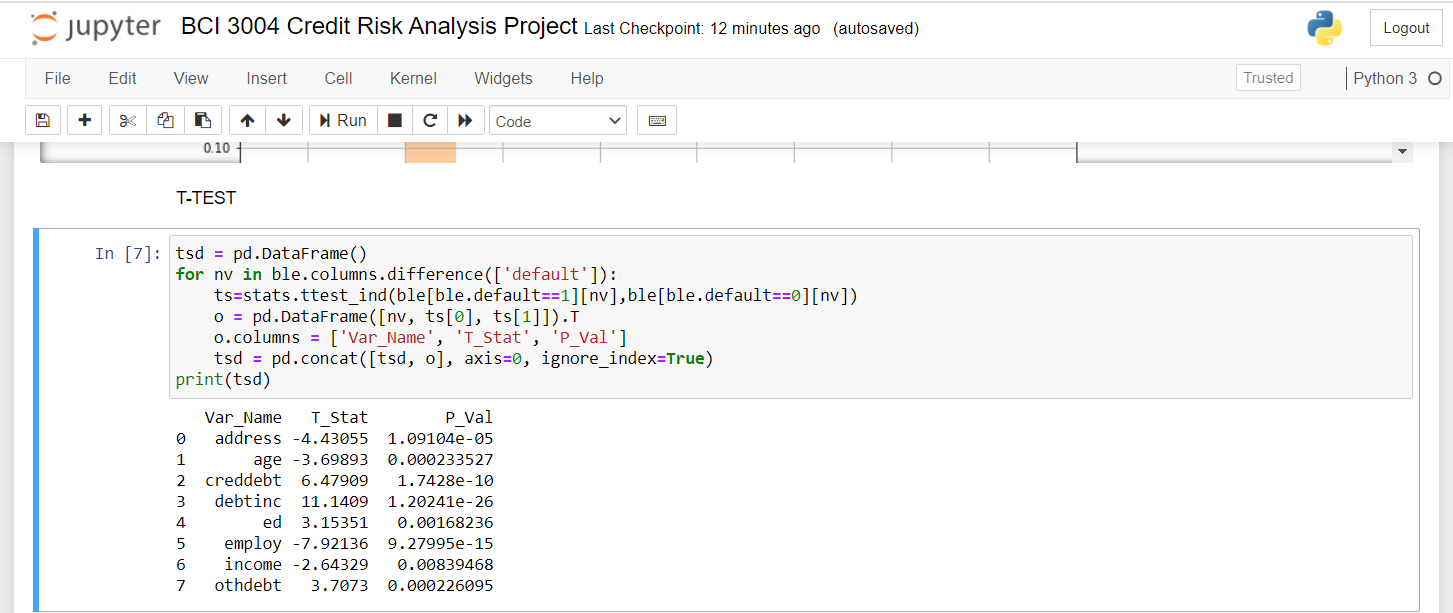


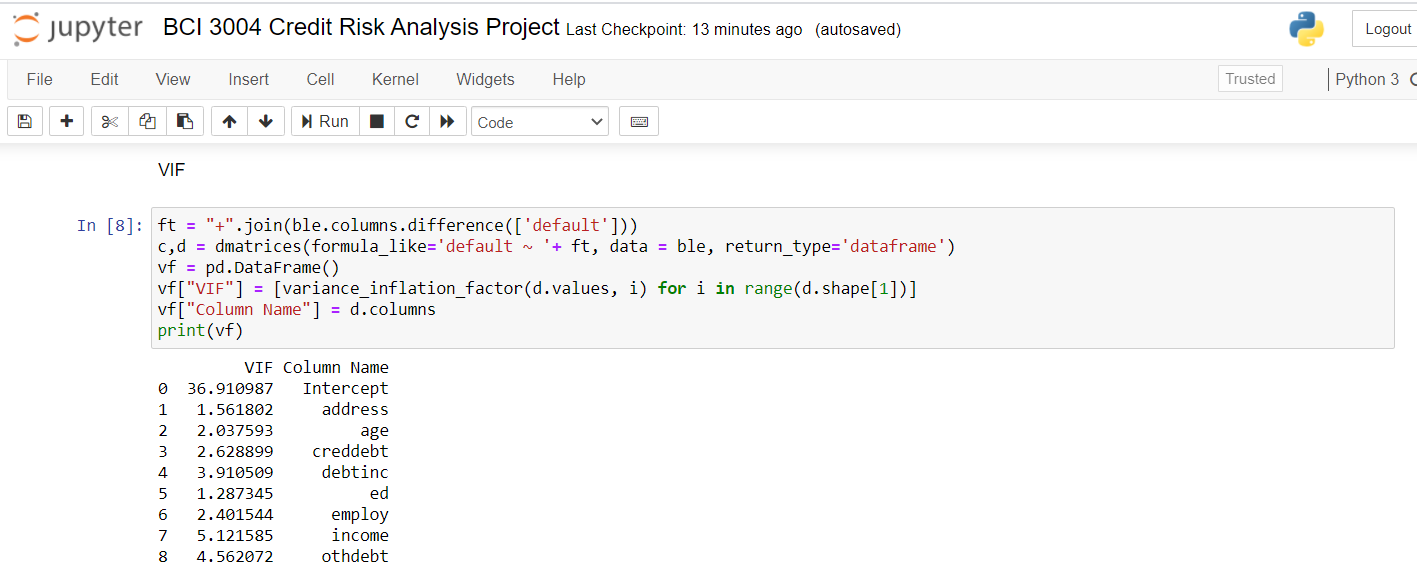




**ii) T-test, VIF**

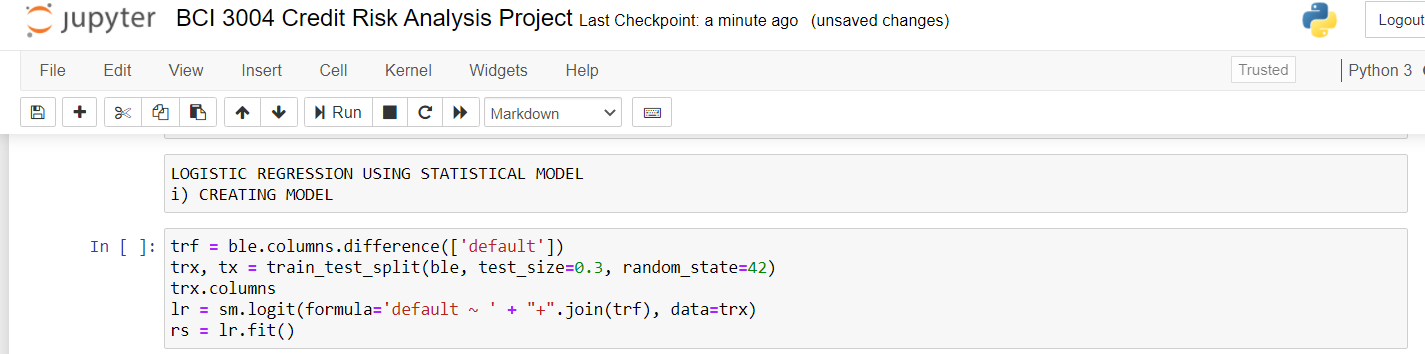




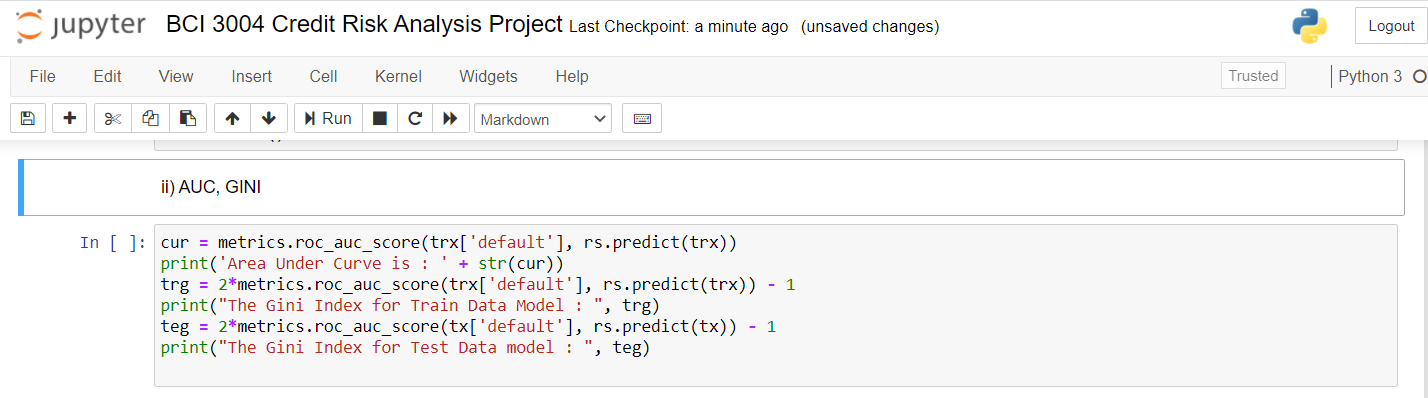


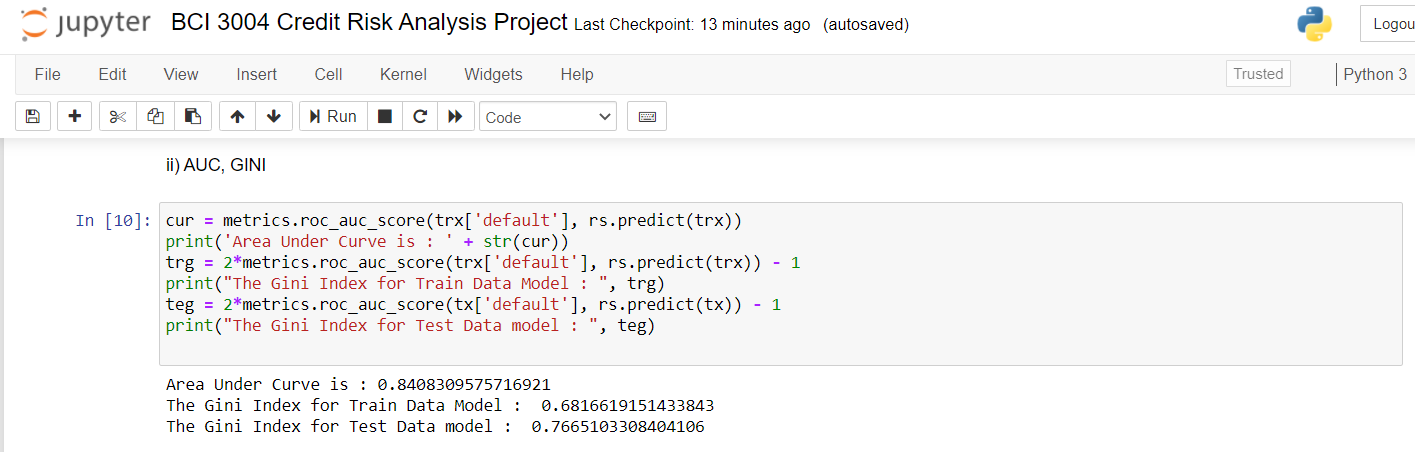
**4. Logistic Regression using Statistical Model**

**i) Creating model**

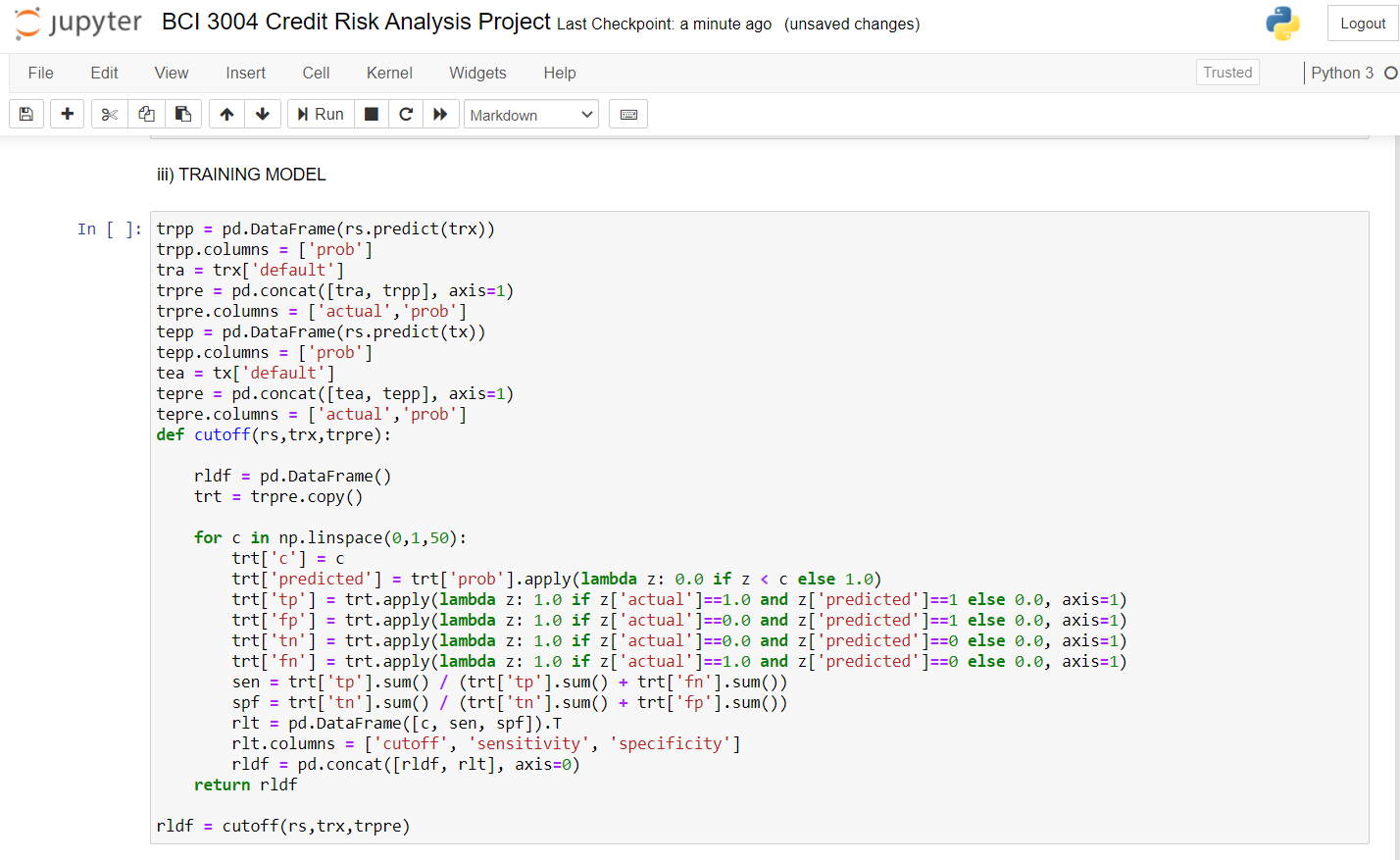


**ii) AUC, Gini Index**

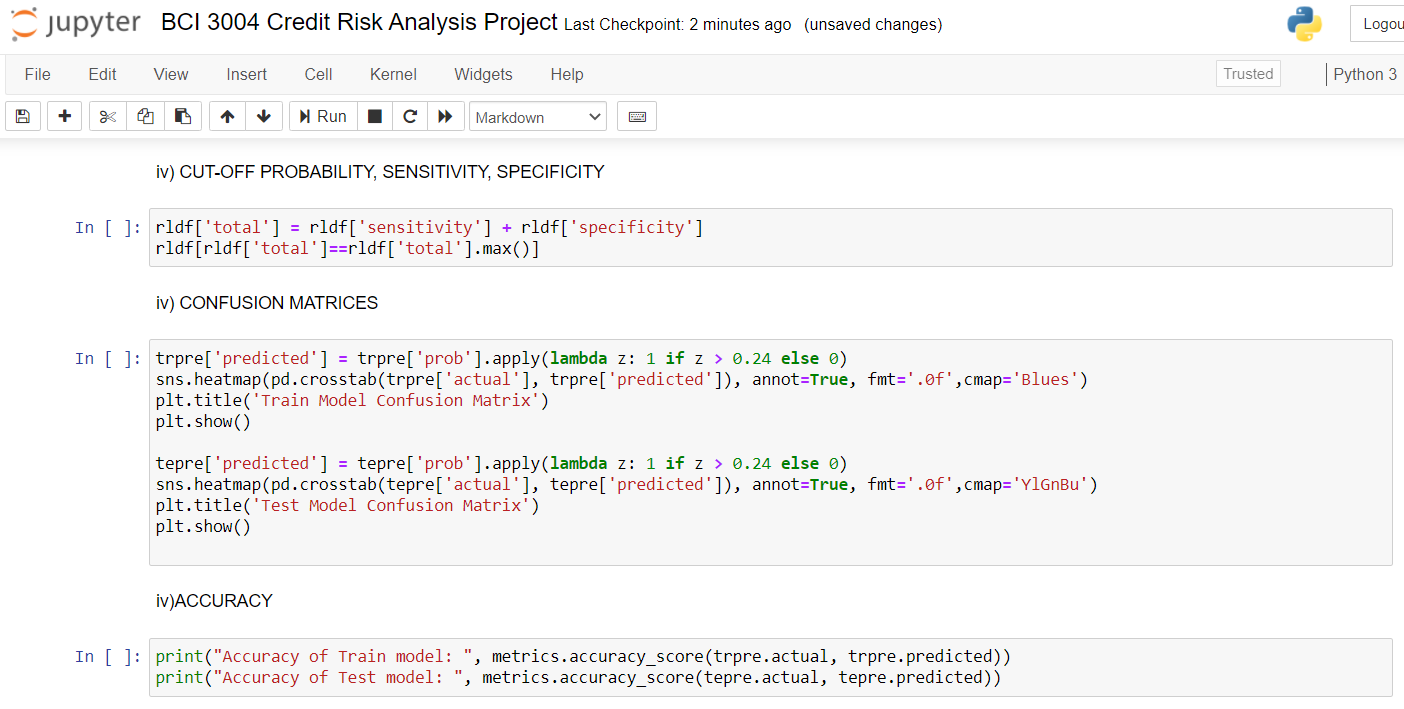


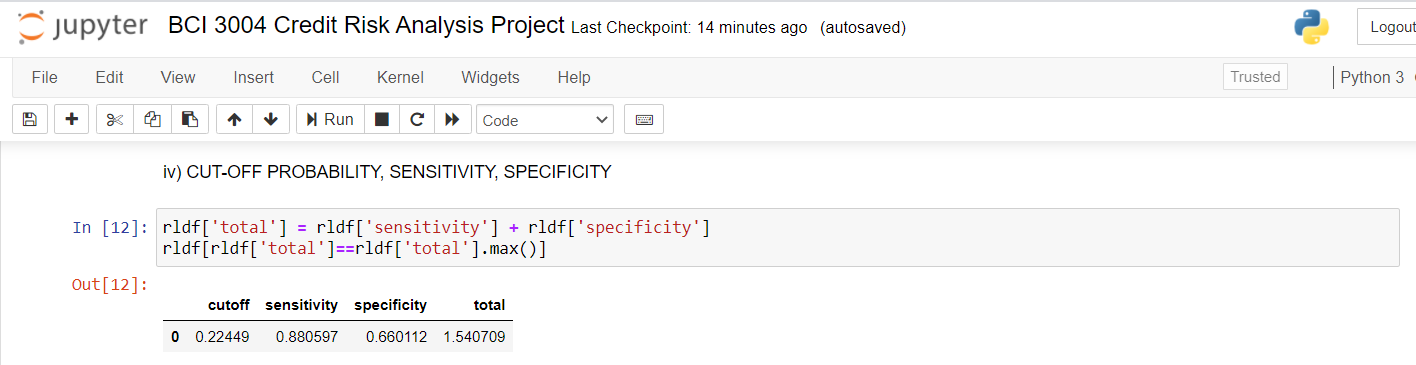


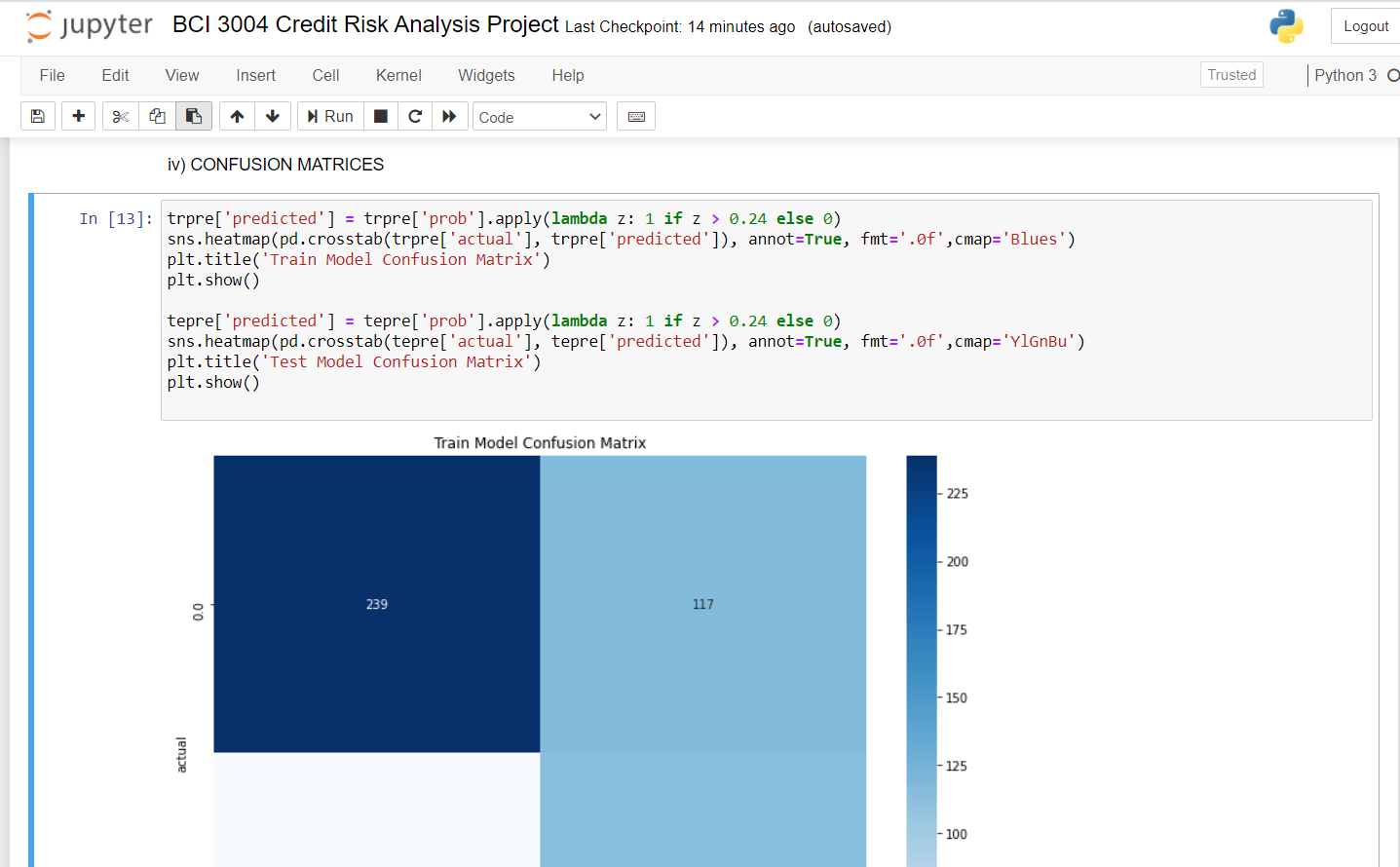
**iii) Training and testing the model**

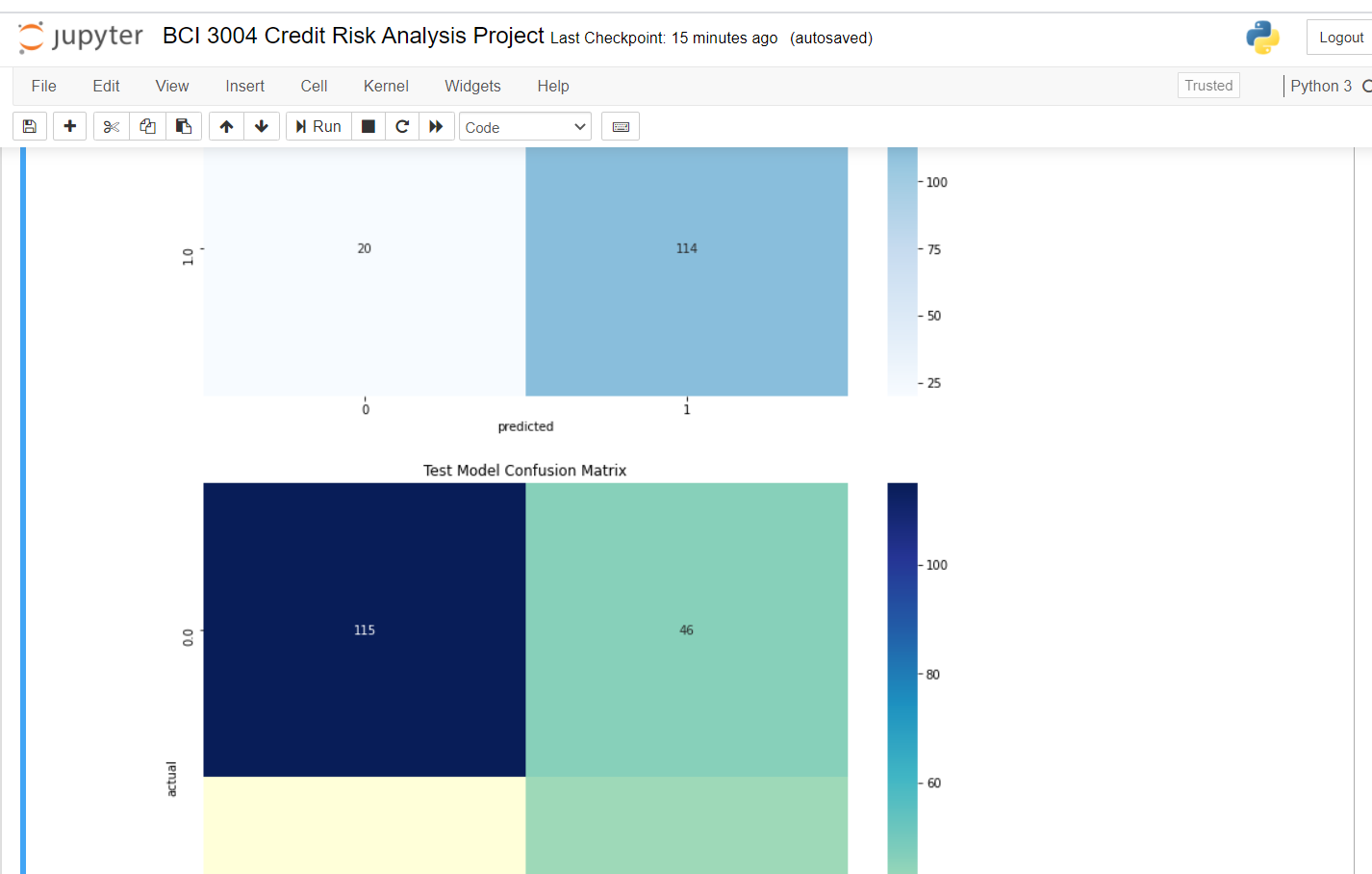


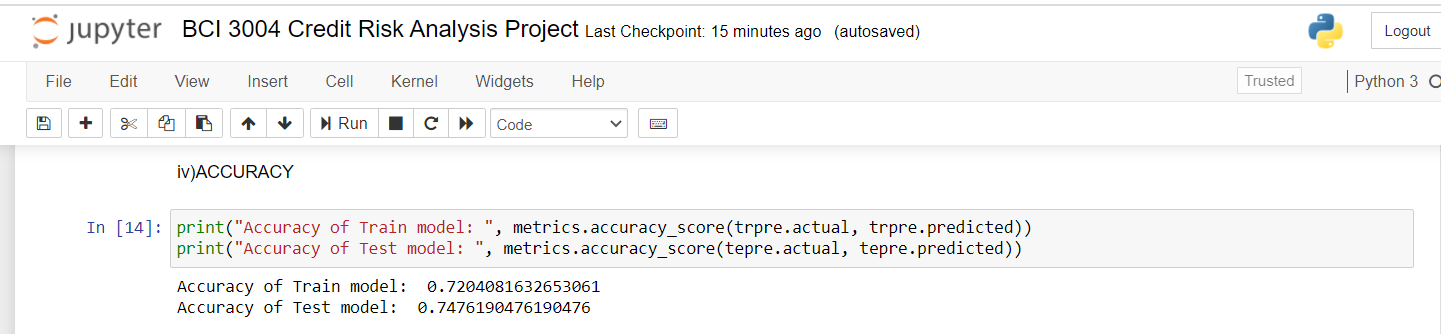
**iv) Cut-off probability, Sensitivity, Specificity, Confusion Matrices, Accuracy**



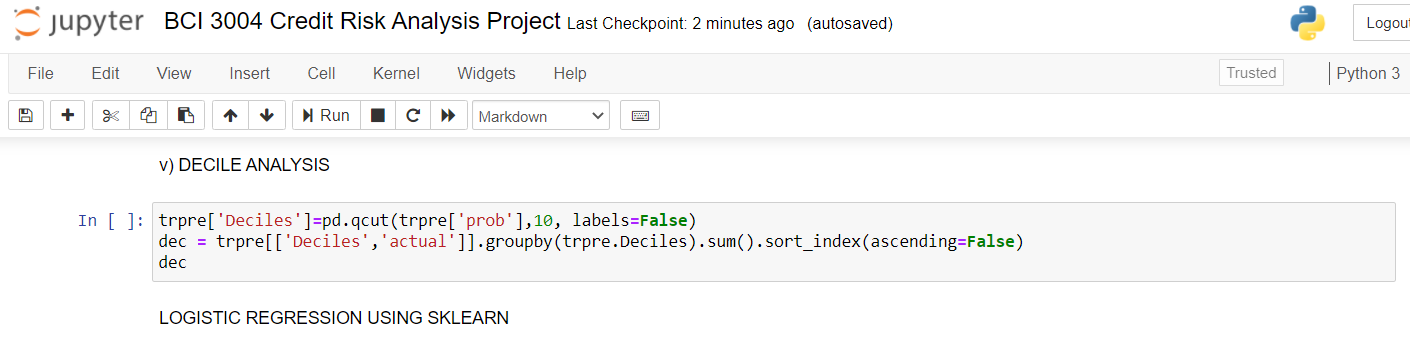


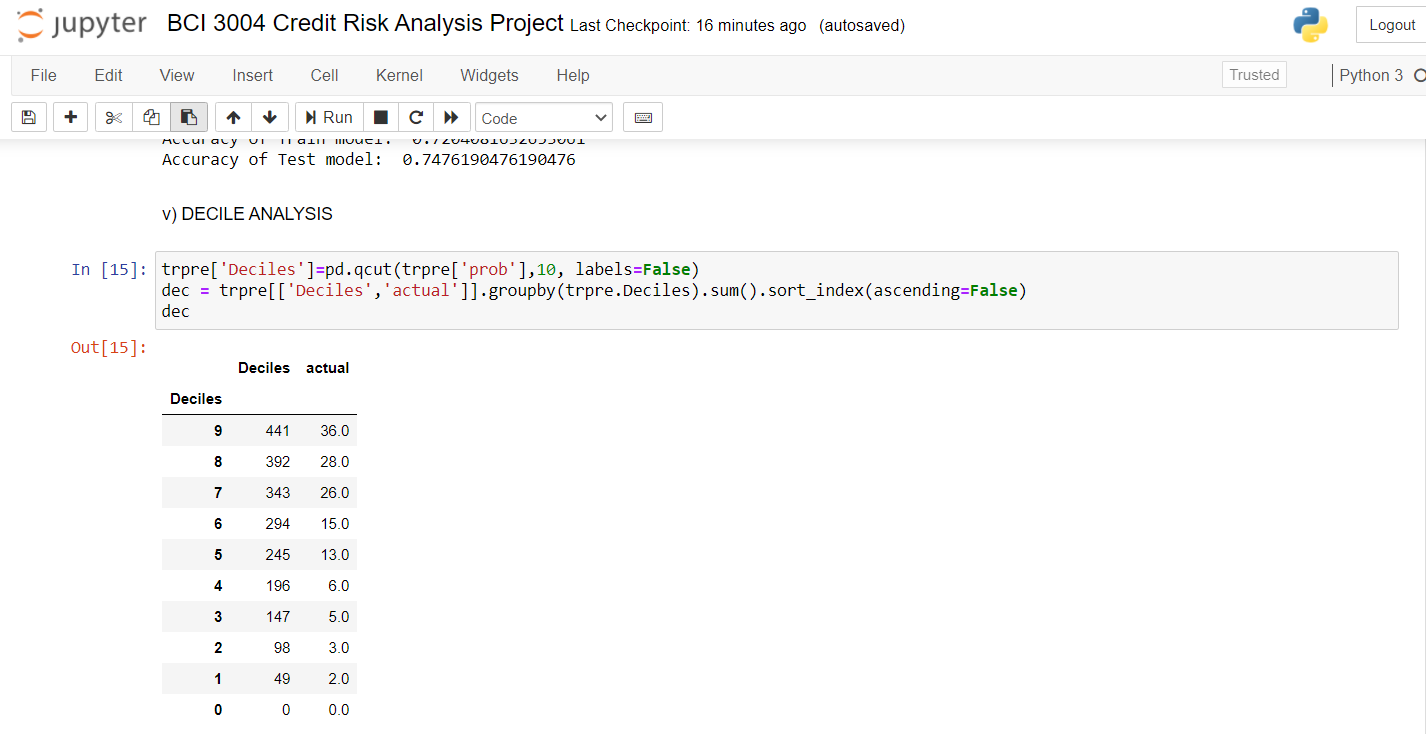






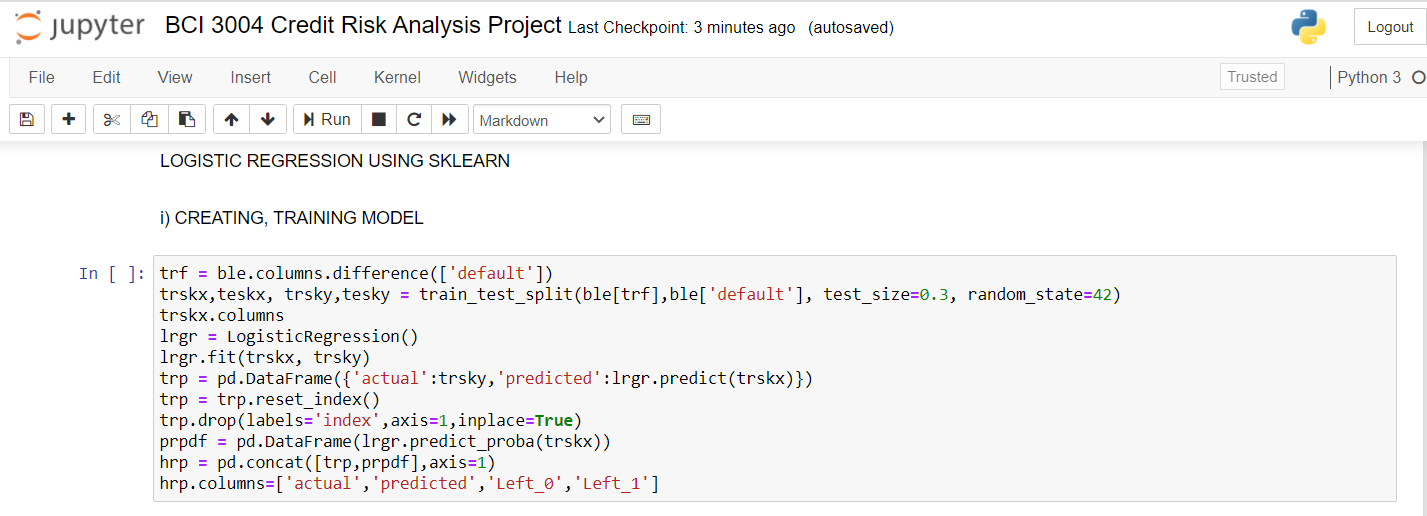
**v) Decile Analysis**



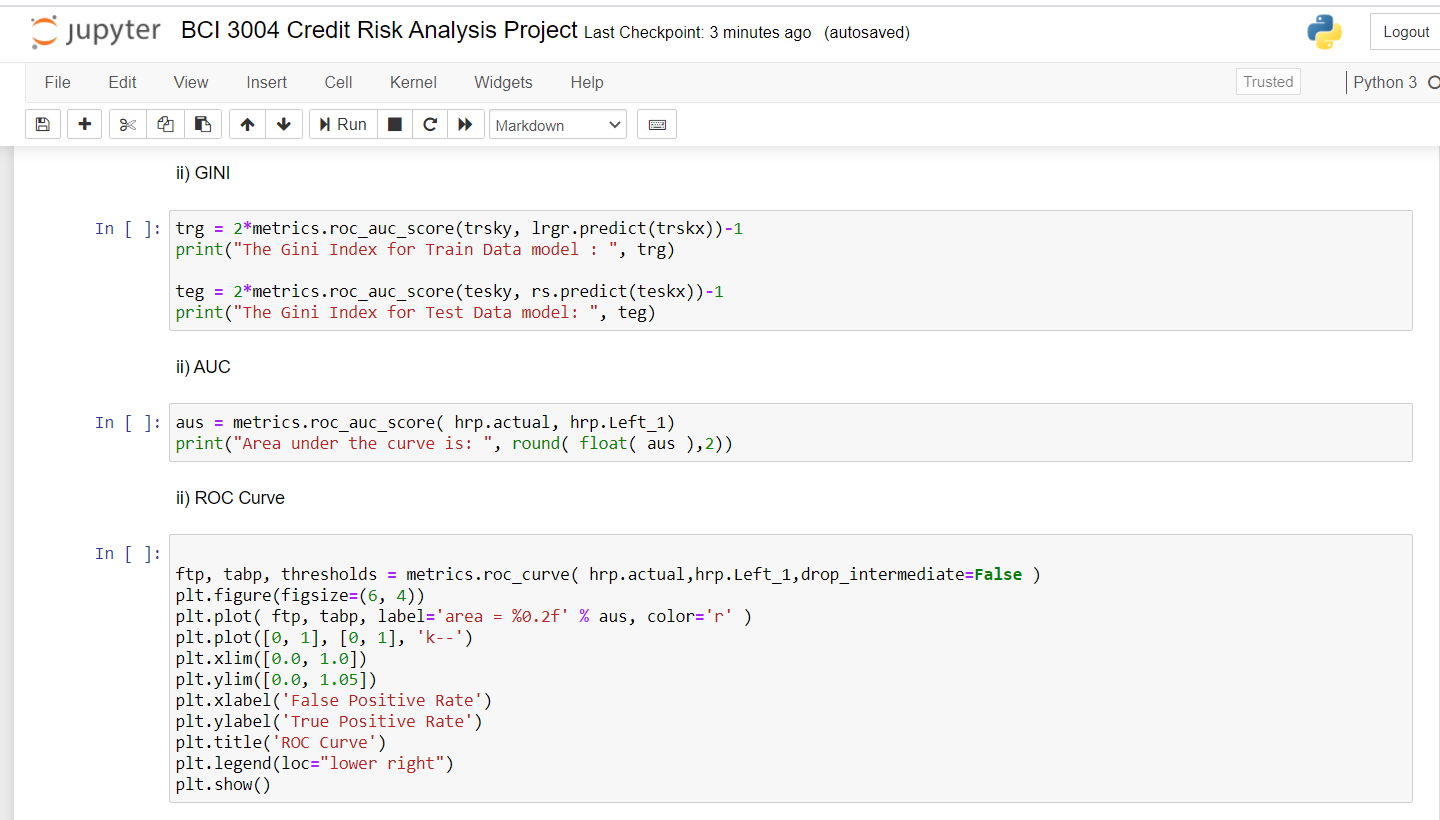


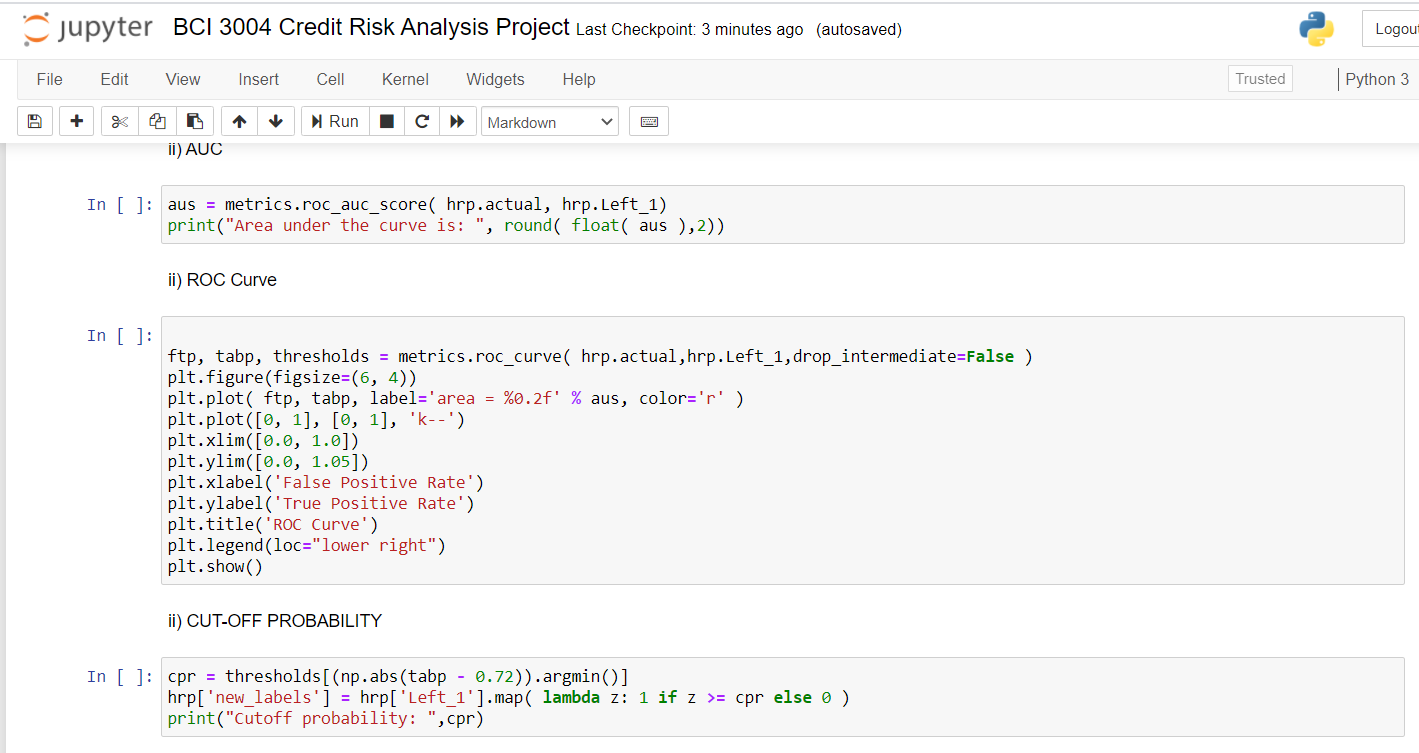
**5. Logistic Regression using sklearn**

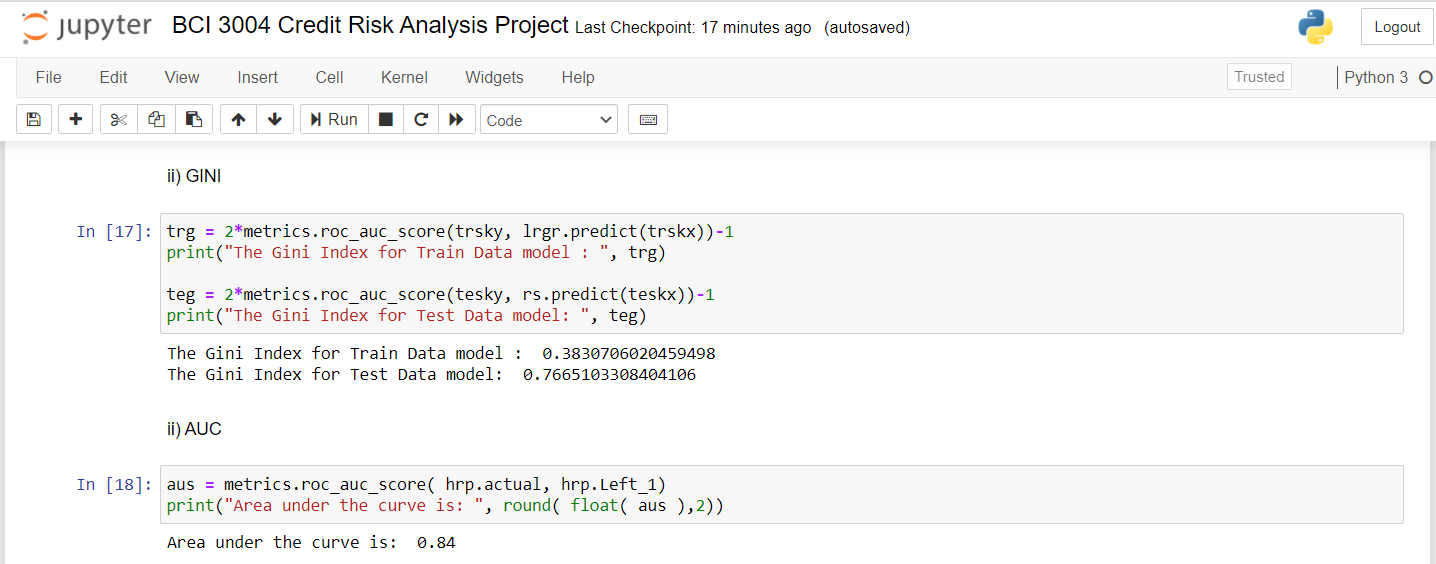
**i) Creating, training model**

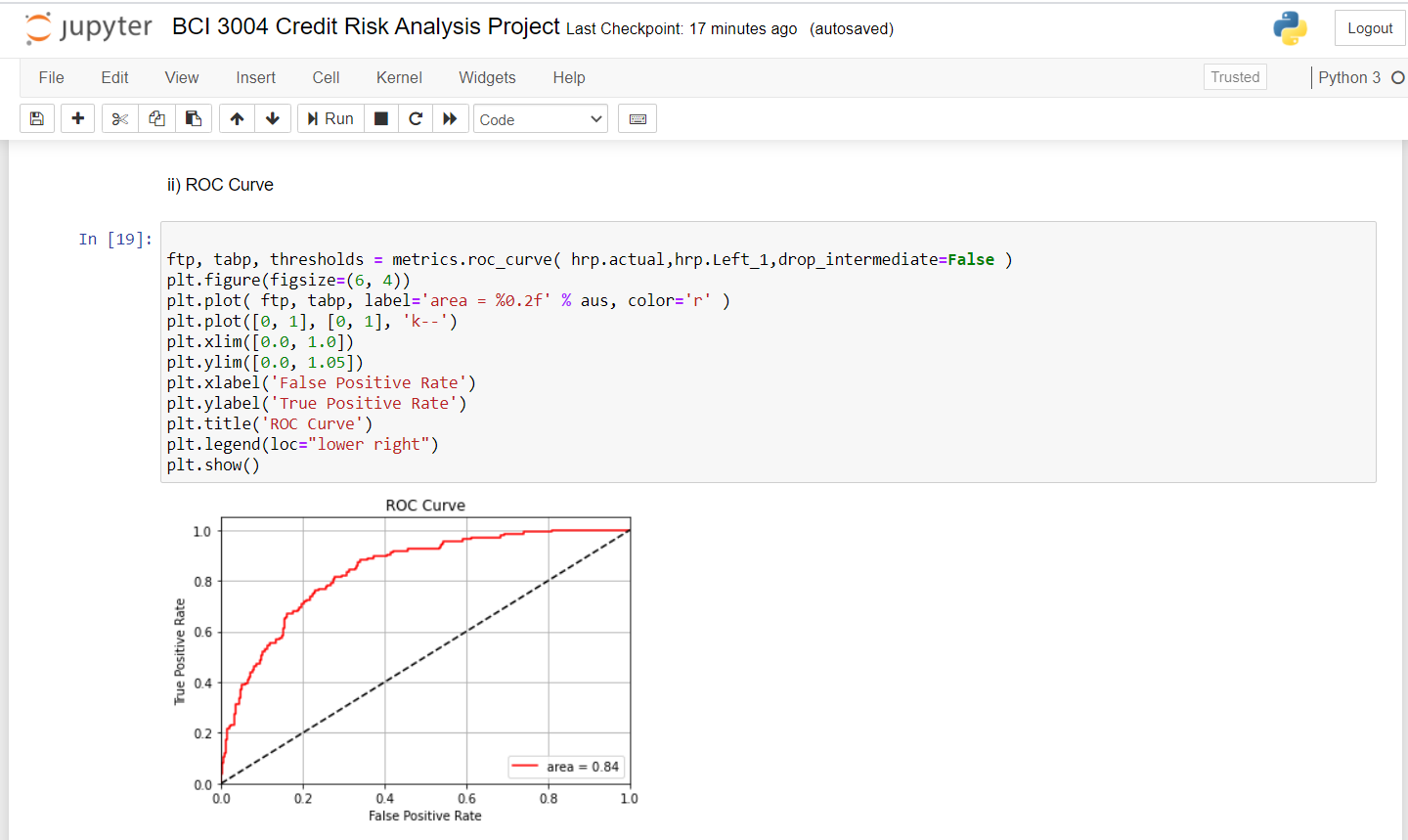


**ii) Gini Index, AUC, ROC Curve, Cut-Off Probability**











**iii) Accuracy of the model using Logistic Regression by sklearn**

