**CREDIT CARD APPROVAL ML MODEL**



Shiva Gowri

Lovely Professional University Department of Computer Science

**ABSTRACT**

This project aims to build a model that can give results on whether a financial institution can approve credit cards to its customer. This card approval decision by financial companies is done based on considering various reasons related to individuals varying from creditworthiness, loan and repayment history, and income standards. This model can help an institution to make a precise judgment on whether a card can be approved or denied for avoiding fraudulence that can impact financial companies with loss. Through the project work, I tried to examine what are the keynote features or requirements considered for issuing a credit card to consumers by financial institutions by evaluating the existing data set Credit Card Approval Dataset from UCI Machine Learning Repository.

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to my advisors and my mentor Dr. Sagar Pande for the continuous support of this study, for motivation, and for sharing their immense knowledge. Their guidance helped me throughout research and writing of this report.

**TABLE OF CONTENTS**

1. ABSTRACT
2. ACKNOWLEDGMENT
3. TABLE OF CONTENTS
4. INTRODUCTION
5. BACKGROUND
6. MACHINE LEARNING
7. PREDICTIVE ANALYSIS
8. ARCHITECTURE
9. IMPLEMENTATION
10. ANALYSIS OF RESULTS
11. MODELS USED:
12. LOGISTIC REGRESSION
13. K-NEAREST NEIGHBORS
14. SUPPORT VECTOR MACHINE
15. DECISION TRESS
16. RANDOM FOREST
17. PROGRAM CODE:
18. CONCLUSION
19. REFERENCES

**INTRODUCTION**

In Current times, everything is completely changed as a digital attribute. One of those digitalized areas is cashless transaction activity. This is very common nowadays, and more people are inclined towards this as this reduces the risk of misplacing cash physically. So, many financial institutions are providing cashless means for their users like debit and credit cards. One of the most prominent options is a credit card. Most people rely on credit cards to perform their transaction activities as it is a very easy way of making their payments. The decisiveness by many financial institutions like national and private banks rely on consumer information like their basic info, living standards, salary, yearly and monthly returns, their current livelihood income source. All this info is reviewed for considering an application. This complete check and analysis can avoid bearing a lot of technical and non-technical losses to the institution. This proper analysis is required as we see tremendous growth in this business sector to avoid any kind of potential risk related to the unethical consumer. precise verification needs to be incorporated by banks when granting credit card to the applicant. Even though decision-making

differs from bank to bank, but the most common factor considered by financial institutions is the consumer’s credit score. As we are seeing an increase in the large growth margin of the credit business of the financial institution due to more consumers interested in applying for credit cards, there is a need to completely automate the process in order to fasten the approval decision by banks. This helps the bank in improving business along with saving time and need of less manpower which is a major saving in terms of money. The model needs to identify the consumers who applied for credit card into two sectors: “No Risk Present” which means the bank can lend money and there is a guarantee that consumer will pay back and banks will not undergo any risk and loss and “risk present” which means banks shouldn’t approve any credit because there is a high chance that consumer can do fraud and banks can undergo financial loss. This classification is done by considering various factors of the consumer like age, salary, the number of years he/she is been working, yearly income, assets, source of income, credit score, repay history, and existed loan dues. These entire mechanisms are not only applicable for a

single consumer, but also to business whether large scale or small scale. In the past, there were various methods introduced to examine the loan history of the consumer and to improve the precision of credit score. These models are data mining models that can be categorized into data that depends on statistical distribution and data which not necessarily depends on the distribution of data. The best example of the model which relies on data distribution is the logistic regression. The linear regression model analysis is used in generating credit scores but this analysis is not favourable since data considered for approving and declining are completely different. Logistic regression supports the data unlike the linear model for parametric tests. Other decision support tools like decision trees, vector machine support is used for non-parametric tests in machine learning. In recent research going on data mining, hybrid approach-based methods are giving optimal results. The neural network approach is considered a better approach to increasing the accuracy of the credit score prediction. This paper proposes a model that predicts whether a credit card can be issued or declined to the applicant by a financial institution.

Even though the decision of the banks is unique and made based on their organization-designed rules but there are certain similar features that considered are similar and those features are taken into consideration when the algorithm is implemented. Data is Credit Card Approval Dataset from UCI Machine Learning Repository. Machine learning pre-processing techniques are implemented and data transformation techniques like scaling, handling of missing values by predefined methods called mean imputation and label encoding for numeric and non-numeric data, dividing data into test and train sets, applying classifiers and to conclude the paper obtained results are further examined using metrics like confusion matrix to examine the accuracy of the result.

**BACKGROUND**

**MACHINE LEARNING**

It is the most rapidly evolving artificial intelligence application. It addresses how a computer or system can execute a task based on a series of text or data without human intervention. This reduces the workload for the people and supports more automation which can save time and money in the long run. We see an enormous amount of data is being tackled by the system in very few seconds and by incorporating machine learning techniques to this fast-paced system can help in building a model. Modelling can perform analysis on high standard and heavy multiplex data and predict the patterns. It gives insight about risk and profits to any business model because machine learning is all about linking and associating data links and predicts relations among them which can be used while taking decisions for the future of the business model. Machine learning algorithms performs its task by recognizing the patterns of the data. The process of identifying patterns is done in two ways: supervised and unsupervised learning. In supervised learning, labelled data is processed by machine learning algorithm to generate the desired trained model. After generating the model, unknown data is used on the model to

generate the result. The well-known machine learning algorithms that use the concept of supervised learning are Random Forest, Linear Regression, Decision Trees, K-nearest Neighbors, Logistic Regression and Support Vector Machines. Since output is obtained based on conclusion drawn from the labelled data, this mechanism can be used to solve real world problems.

In unsupervised Learning, unlabelled data is processed by machine learning algorithm to generate trained model. Since data is unknown, the trained model will identify patterns of the data. It helps in categorization by identifying prominent features. Clustering and association are mechanisms opted in unsupervised learning. Large scale industries that need to handle high volumes of data more efficiently like Financial Service industry, Health Care, Retail, Government and Transportation uses this technology to process high amount of data to obtain predictions

Classification Algorithm in Machine Learning

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

y=f(x), where y = categorical output

The best example of an ML classification algorithm is **Email Spam Detector**.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.

The algorithm which implements the classification on a dataset is known as a classifier. There are two types of Classifications:

* **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called as Binary Classifier.  
  **Examples:** YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.
* **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.  
  **Example:** Classifications of types of crops, Classification of types of music.

**PREDICTIVE ANALYSIS**

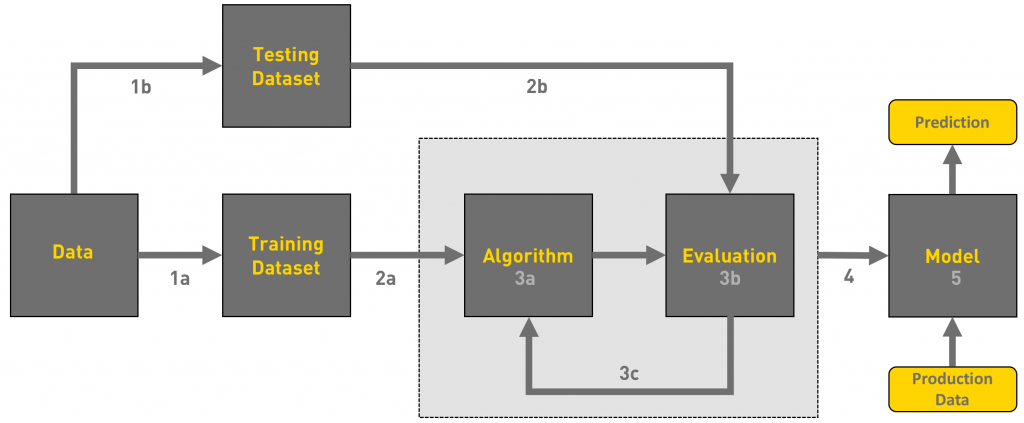
When we do predictive data mining, we analyse the work based on the observation that is drawn from the available existing data along with the use of new or added factors to identify hidden patterns that are interlinked to rule out conclusions.

**ARCHITECTURE**

The implementation of the project is done in multiple steps by applying various techniques. The steps vary from analysis of dataset by observing, processing the data by identifying anomalies or data that is needed to be converted since dataset available can be masked for various security reasons. Further, handling the missing values in the dataset taken, then dividing data into two sets such that one set is used for training so that we can develop the model and another set is used for testing and verifying the model for accuracy

**The flow chat shows us the working of the entire model**





**IMPLEMENTATION**

The first step in any analysis is to obtain the [dataset](http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/) and [codebook](http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.names). Both the dataset and the codebook can be downloaded for free from the UCI website. A quick review of the codebook shows that all of the values in the dataset have been converted to meaningless symbols to protect the confidentiality of the data. This will still suit our purposes as a demonstration dataset since we are not using the data to develop actual credit screening criteria. However, to make it easier to work with the dataset, I gave the variables working names based on the type of data.

Once the dataset is loaded, we’ll use the info() function to quickly understand the type of data in the dataset. This function only shows the first few values for each column so there may be surprises deeper in the data but it’s a good start. Here you can see the names assigned to the variables. The first 15 variables are the credit application attributes. The Approved variable is the credit approval status and target value.

Using the output below, we can see that the outcome values in Approved are ‘+’ or ‘-’ for whether credit had been granted or not. These character symbols aren’t meaningful as is so will need to be transformed. Turning the ‘+’ to a ‘1’ and the ‘-’ to a ‘0’ will help with classification and logistic regression models later in the analysis. To perform data visualization analysis, a software library of python called pandas is used.

'Data. frame': 690 obs. of 16 variables:

$ Male : num 1 1 0 0 0 0 1 0 0 0 ...

$ Age : chr "58.67" "24.50" "27.83" "20.17" ...

$ Debt : num 4.46 0.5 1.54 5.62 4 ...

$ Married : chr "u" "u" "u" "u" ...

$ Bank Customer : chr "g" "g" "g" "g" ...

$ Education Level: chr "q" "q" "w" "w" ...

$ Ethnicity : chr "h" "h" "v" "v" ...

$ Years Employed : num 3.04 1.5 3.75 1.71 2.5 ...

$ Prior Default : num 1 1 1 1 1 1 1 1 1 0 ...

$ Employed : num 1 0 1 0 0 0 0 0 0 0 ...

$ Credit Score : num 6 0 5 0 0 0 0 0 0 0 ...

$ Driver’s License: chr "f" "f" "t" "f" ...

$ Citizen : chr "g" "g" "g" "s" ...

$ Zip Code : chr "00043" "00280" "00100" "00120" ...

$ Income : num 560 824 3 0 0 ...

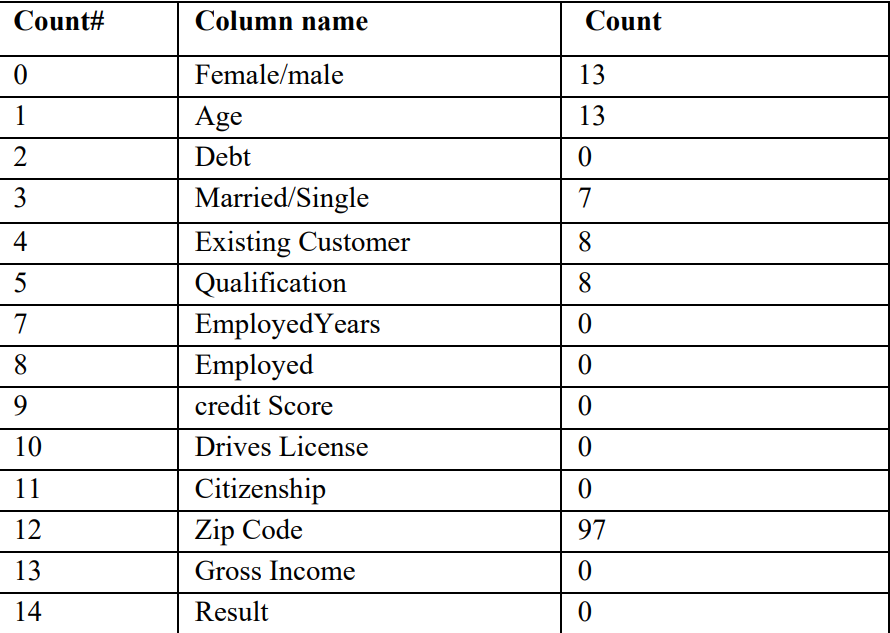
$ Approved : chr "+" "+" "+" "+" ...

the dataset consists of alphanumeric data. Age, debt, years employed and credit score columns in the dataset contain numeric values (int and float) and the rest of the columns are object type.

The dataset contains 590 rows and 16 columns of data. There are some existing missing values that cannot be ignored because they can affect the accuracy of the model. We overwrite question marks with NAN by implementing the replace method from the NumPy library.

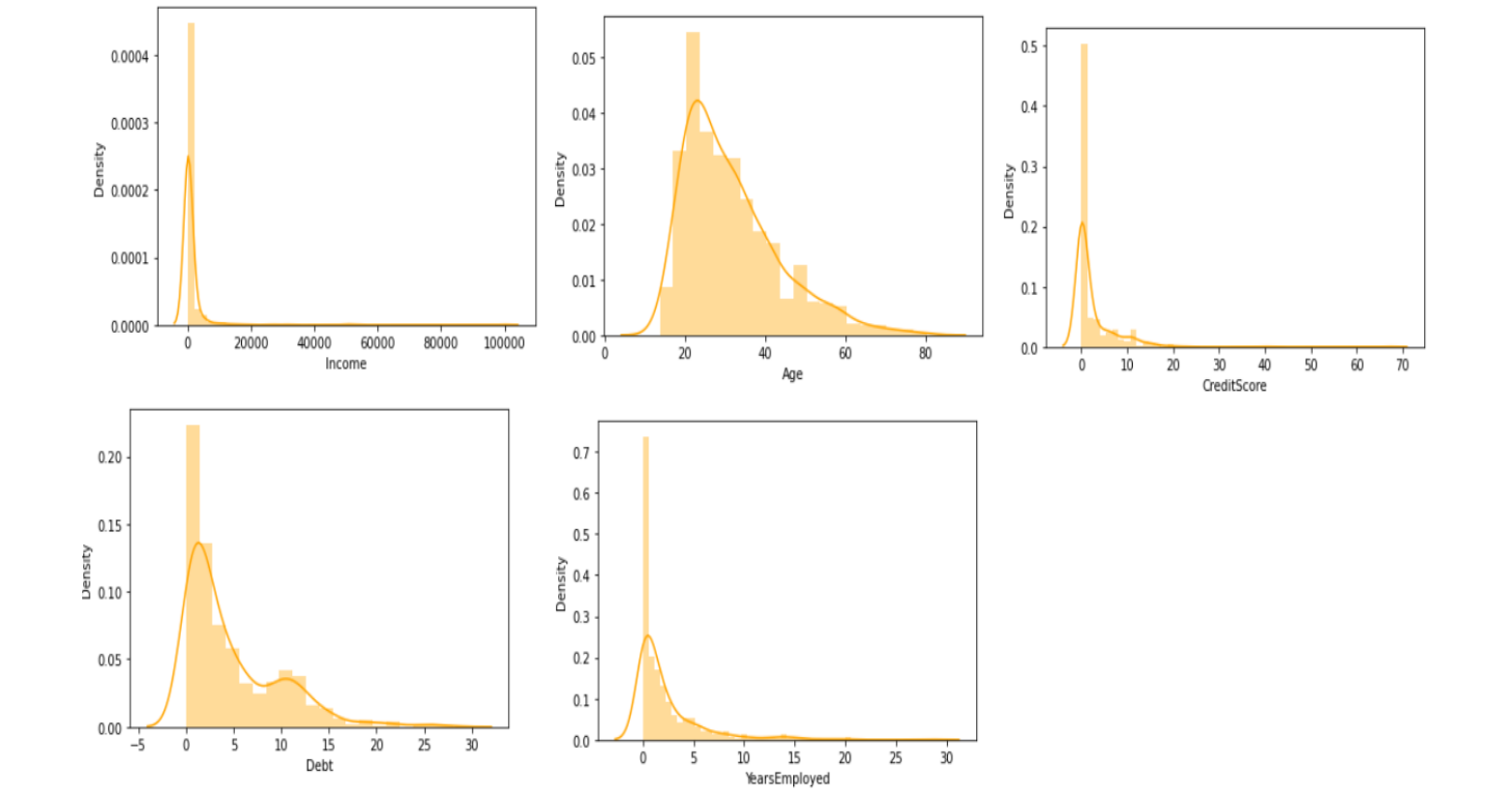
The method called “mean imputation” is performed on columns with only numerical data and replaced NAN with the mean value. Mean imputation is applied over the dataset so that data can be converted or replaced with existing variables mean value.

Zero in the column defines there are no missing values present in those corresponding rows and a number other than zero indicates the presence of missing values. If the row values are numeric, the mean imputation mechanism is implemented and if the row values are non-numeric (object type) then the missing column value will be replaced with the most frequent occurred value in that column of the dataset.

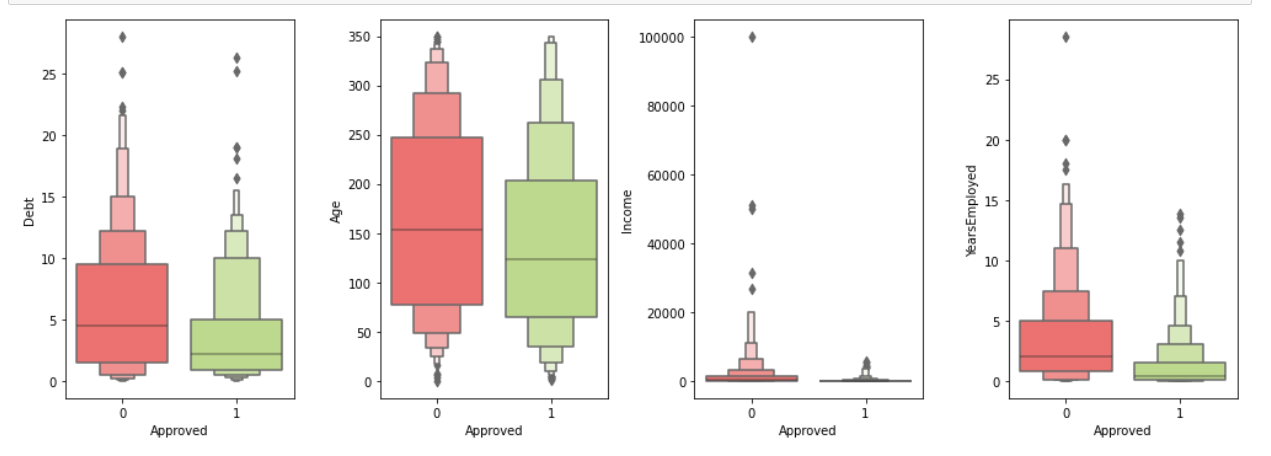


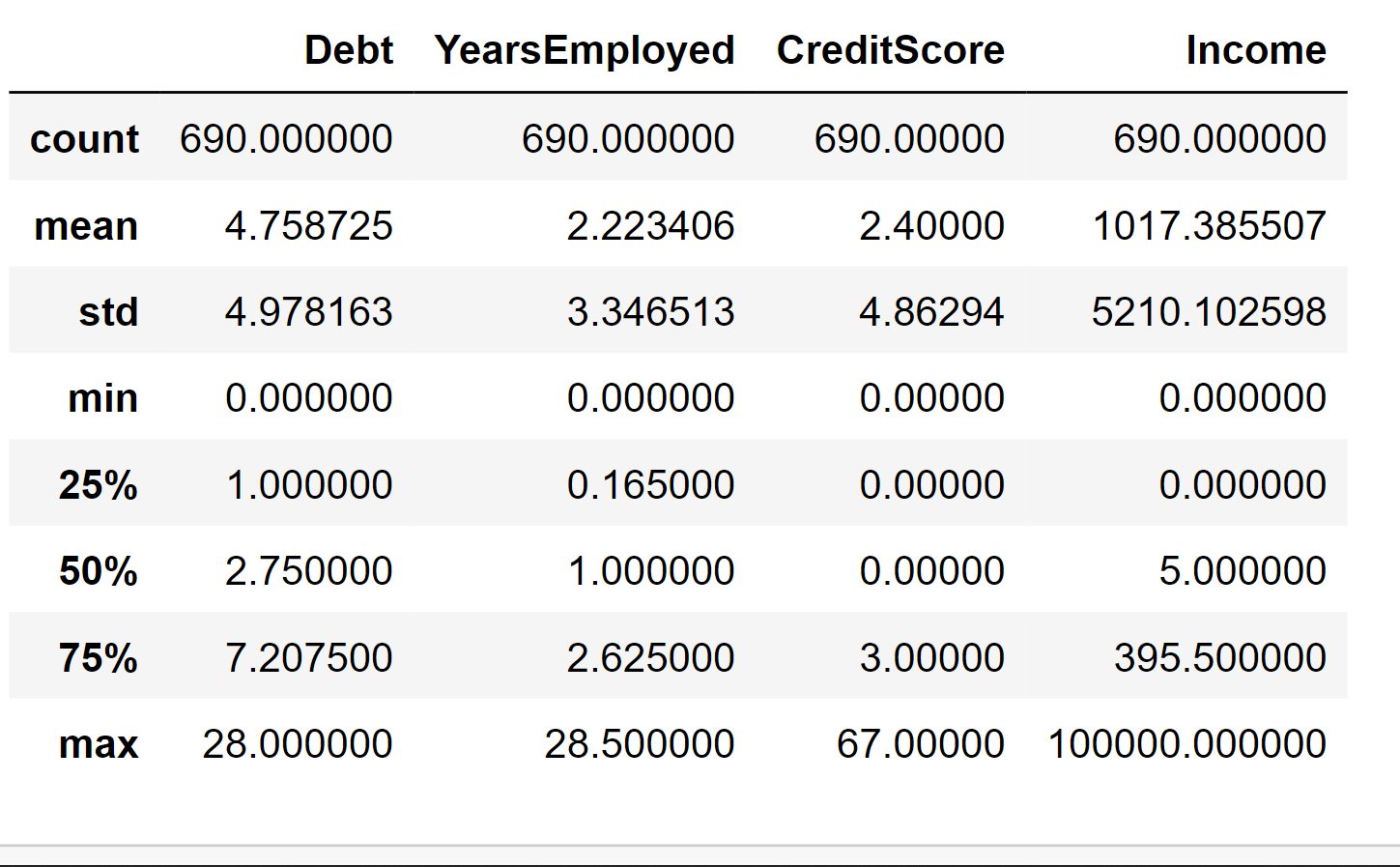
EDA(exploratory data analysis)

The exploratory data analysis approach is opted for categorizing features of numeric values to visualize the actual structure of data. Figure 5 histograms show the distribution of data along with the mean.



Now, we will use the summary () function to see the descriptive statistics of the numeric values such as min, max, mean, and median. The range is the difference between the minimum and maximum values and can be calculated from the summary() output. For the B variable, the range is 66.5 and the standard deviation is 11.9667.





We can see from the summary output that the Debt variable has missing values that we’ll have to fill in. We could simply use the mean of all the existing values to do so. Another method would be to check the relationship among the numeric values and use a linear regression to fill them in. The table below shows the correlation between all of the variables. The diagonal correlation values equal 1.000 because each variable is perfectly correlated with itself. To read the table, we will look at the data by rows. The largest value in the first row is 0.396 meaning age is most closely correlated with YearsEmployed. Similarly, Debt is mostly correlated with YearsEmployed.

Age Debt YearsEmployed CreditScore Income

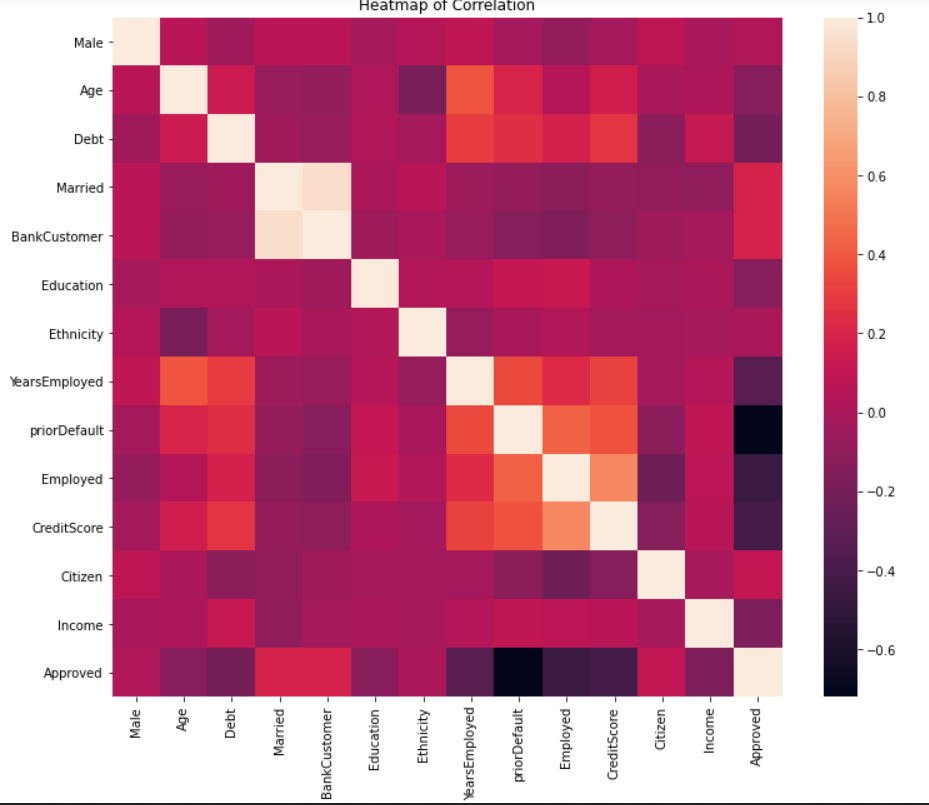
Age 1.000 0.202 0.396 0.186 0.019

Debt 0.202 1.000 0.301 0.271 0.122

YearsEmployed 0.396 0.301 1.000 0.327 0.053

CreditScore 0.186 0.271 0.327 1.000 0.063

Income 0.019 0.122 0.053 0.063 1.000

**HeatMap Of Correlation**

Splitting the dataset into training and test sets. We have successfully converted all the non-numeric values to numeric ones.

Now, we will split our data into train set and test set to prepare our data for two different phases of machine learning modelling: training and testing. Ideally, no information from the test data should be used to scale the training data or should be used to direct the training process of a machine learning model. Hence, we first split the data and then apply the scaling.

Also, features like DriversLicense and ZipCode are not as important as the other features in the dataset for predicting credit card approvals. We should drop them to design our machine learning model with the best set of features. In Data Science literature, this is often referred to as feature selection.

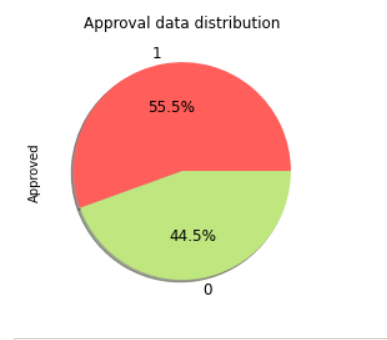
**Analysis Of Results**

The dataset repository contains a mix of both approved and non-approved 690 records. (Pie chart) gives the total number of approved and rejected records in the dataset.

the percentage of both approved (55.5%) and denied (44.5%) records in the dataset.

Approved: 307

Rejected: 383



The reliability of the model can be observed by using a table called the confusion matrix. It is used to visualize the performance of an algorithm. This matrix table is a two-dimensional matrix that can give us a result about true or correct predictions and incorrect or false predictions. The data is 15 divided into two sets which we call test and train sets. Now we use test data and implement the model by using the confusion matrix of the sklearn library and get an accuracy percentage report.

Accuracy is the ratio of the number of correct predictions to the total number of predictions. Code depicts the logistic regression, svm, knn, decision trees, random tress(ensemble) model accuracy and confusion matrix values. In the confusion matrix array, the first element of the first row depicts the number of denied applications predicted correctly and the rightmost element in the second row depicts the number of approved applications predicted correctly

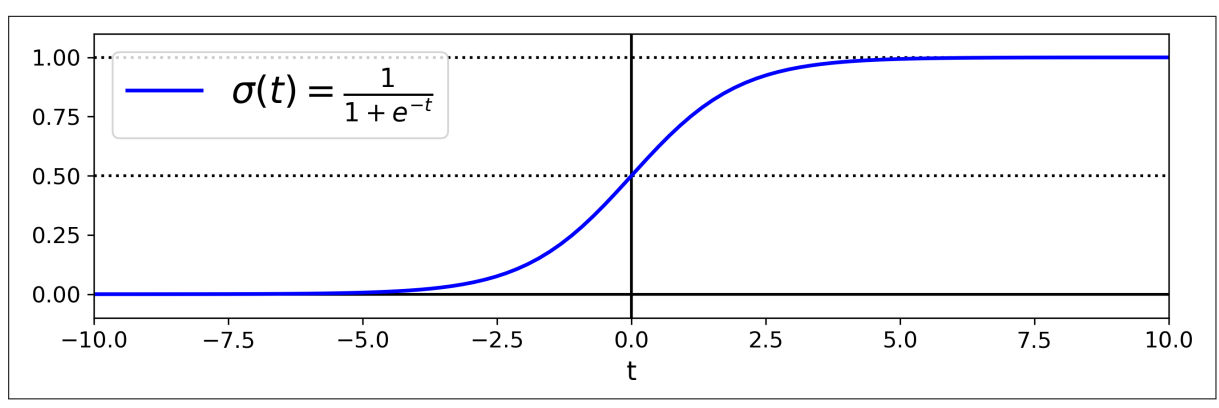
The data is now split into two separate sets — train and test sets respectively. We are only left with one final pre-processing step of scaling before we can fit a machine learning model to the data.

Now, let’s try to understand what these scaled values mean in the real world. Let’s use CreditScore as an example. The credit score of a person is their creditworthiness based on their credit history. The higher this number, the more financially trustworthy a person is considered to be.

**LOGISTIC REGRESSION MODEL**

Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a particular class (e.g., what is the probability that this email is spam?). If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labelled “1”), or else it predicts that it does not (i.e., it belongs to the negative class, labeled “0”). This makes it a binary classifier.

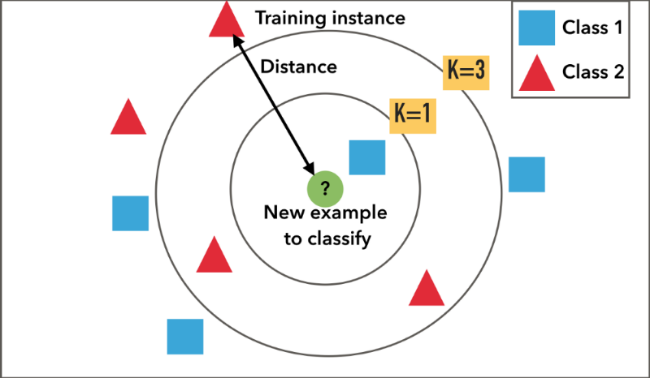
Logistic Regression model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like the Linear Regression model does, it outputs the logistic of this result



**K-NEAREST NEIGHBORS**

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).



**SUPPORT VECTOR MACHINE**

A Support Vector Machine (SVM) is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detection. It is one of the most popular models in Machine Learning, and any‐ one interested in Machine Learning should have it in their toolbox. SVMs are particularly well suited for classification of complex but small- or medium-sized datasets.

The LinearSVC class regularizes the bias term, so you should center the training set first by subtracting its mean. This is automatic if you scale the data using the StandardScaler. Moreover, make sure you set the loss hyperparameter to "hinge", as it is not the default value. Finally, for better performance you should set the dual hyperparameter to False, unless there are more features than training instances



**DECISION TREE CLASSIFIER**

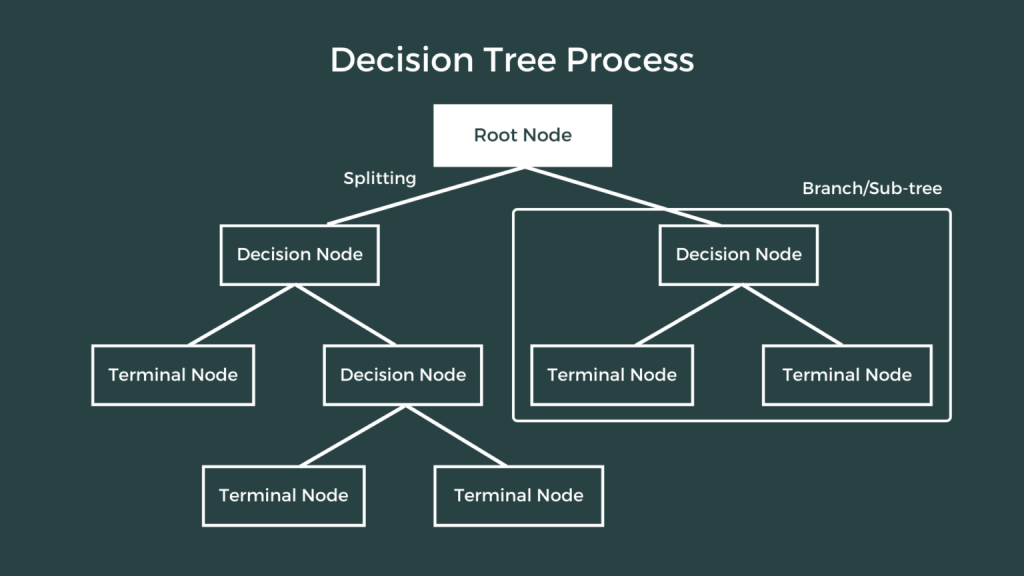
Decision tree classifiers provide a readable [classification model](https://www.sciencedirect.com/topics/computer-science/classification-models) that is potentially accurate in many different application contexts, including energy-based applications. The decision tree classifier (Pang-Ning et al., 2006) creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

Each leaf represents class labels associated with the instance. Instances in the training set are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Starting from the root node of the tree, each node splits the instance space into two or more sub-spaces according to an attribute test condition. Then moving down the tree branch corresponding to the value of the attribute, a new node is created.

This process is then repeated for the subtree rooted at the new node, until all records in the training set have been classified. The decision tree construction process usually works in a top-down manner,

by choosing an attribute test condition at each step that best splits the records. There are many measures that can be used to determine the best way to split the records.

The Gini index impurity-based criterion for growing the tree (Pang-Ning et al., 2006) is often exploited. It measures how often a randomly chosen instance from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.



**RANDOM FORESTS CLASSIFIER**

In random forests each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set.

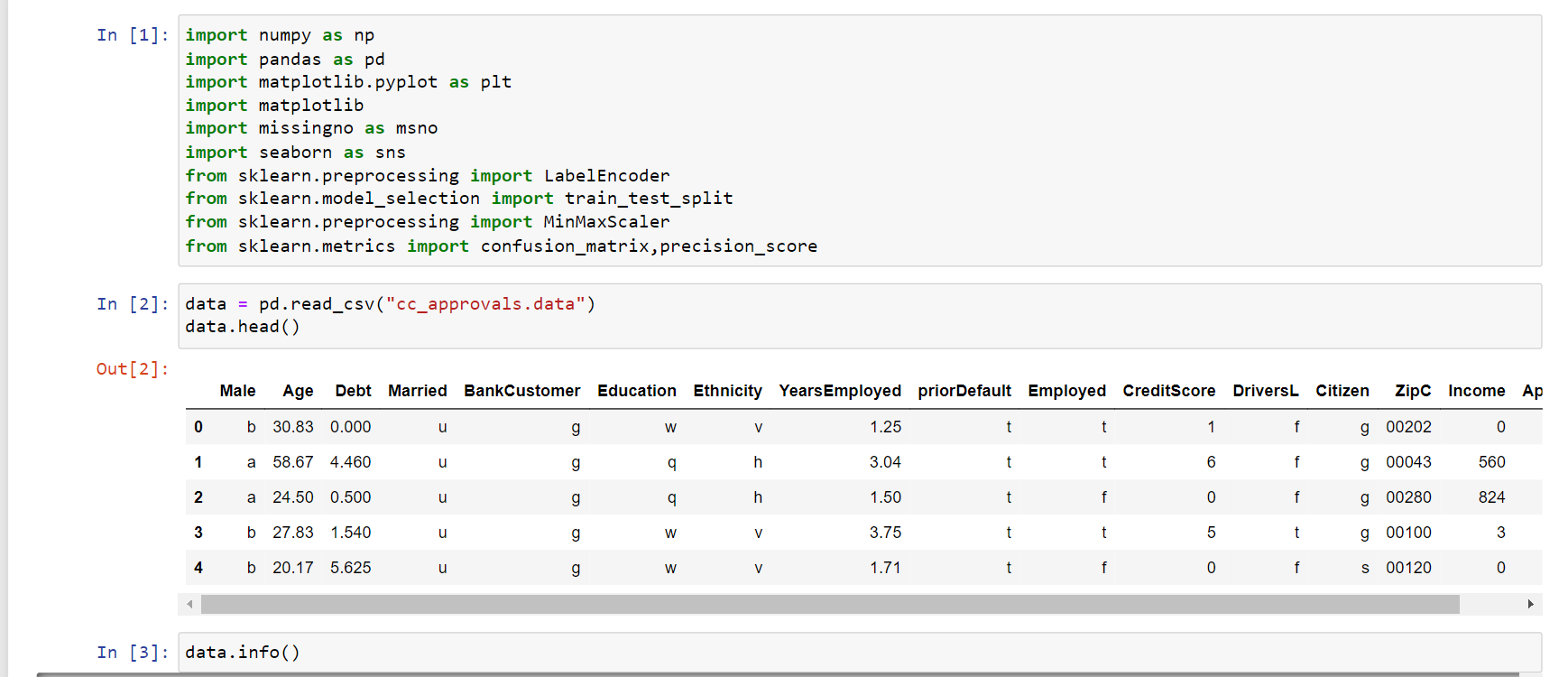
Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max\_features.

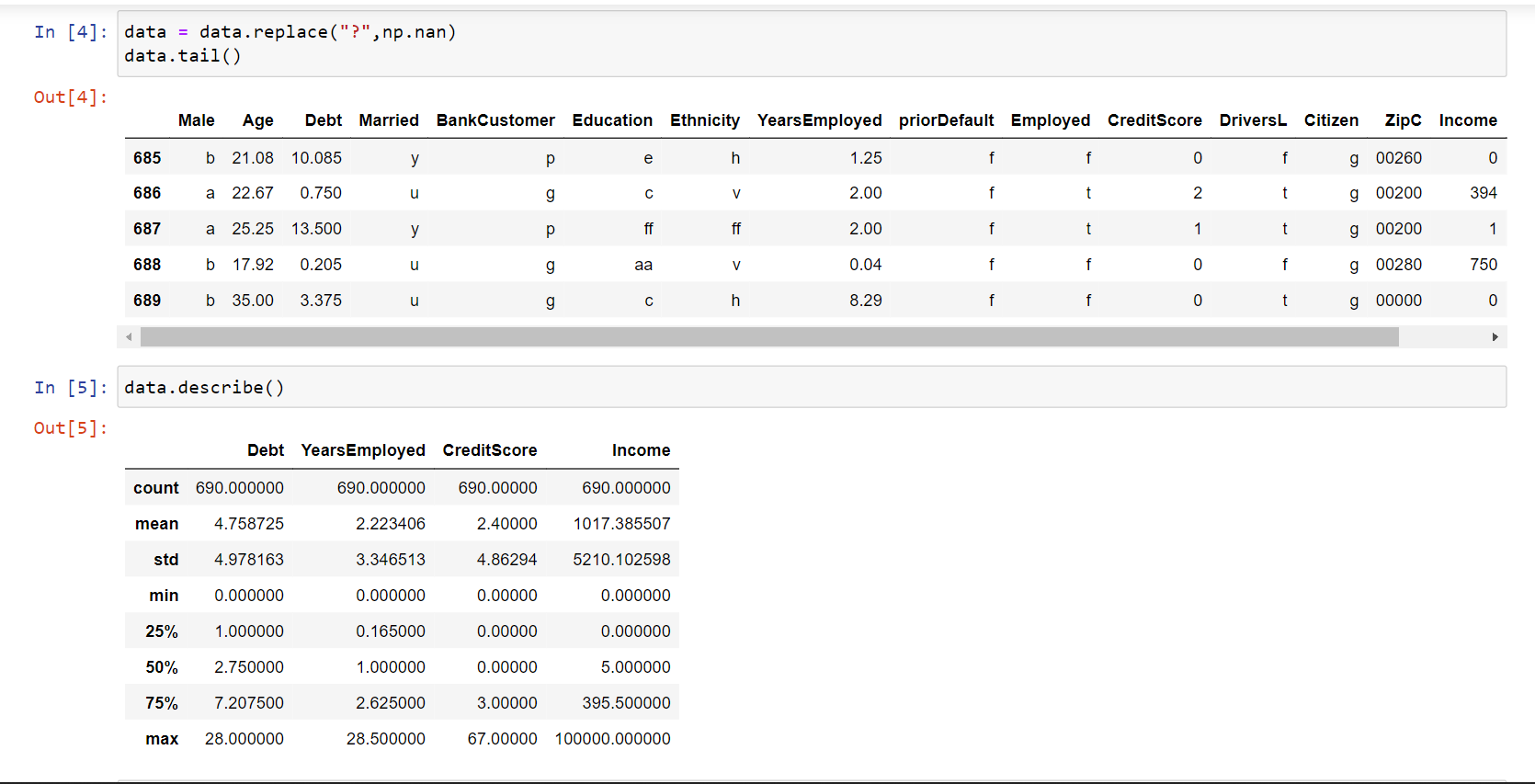
The purpose of these two sources of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yield decision trees with somewhat decoupled prediction errors.

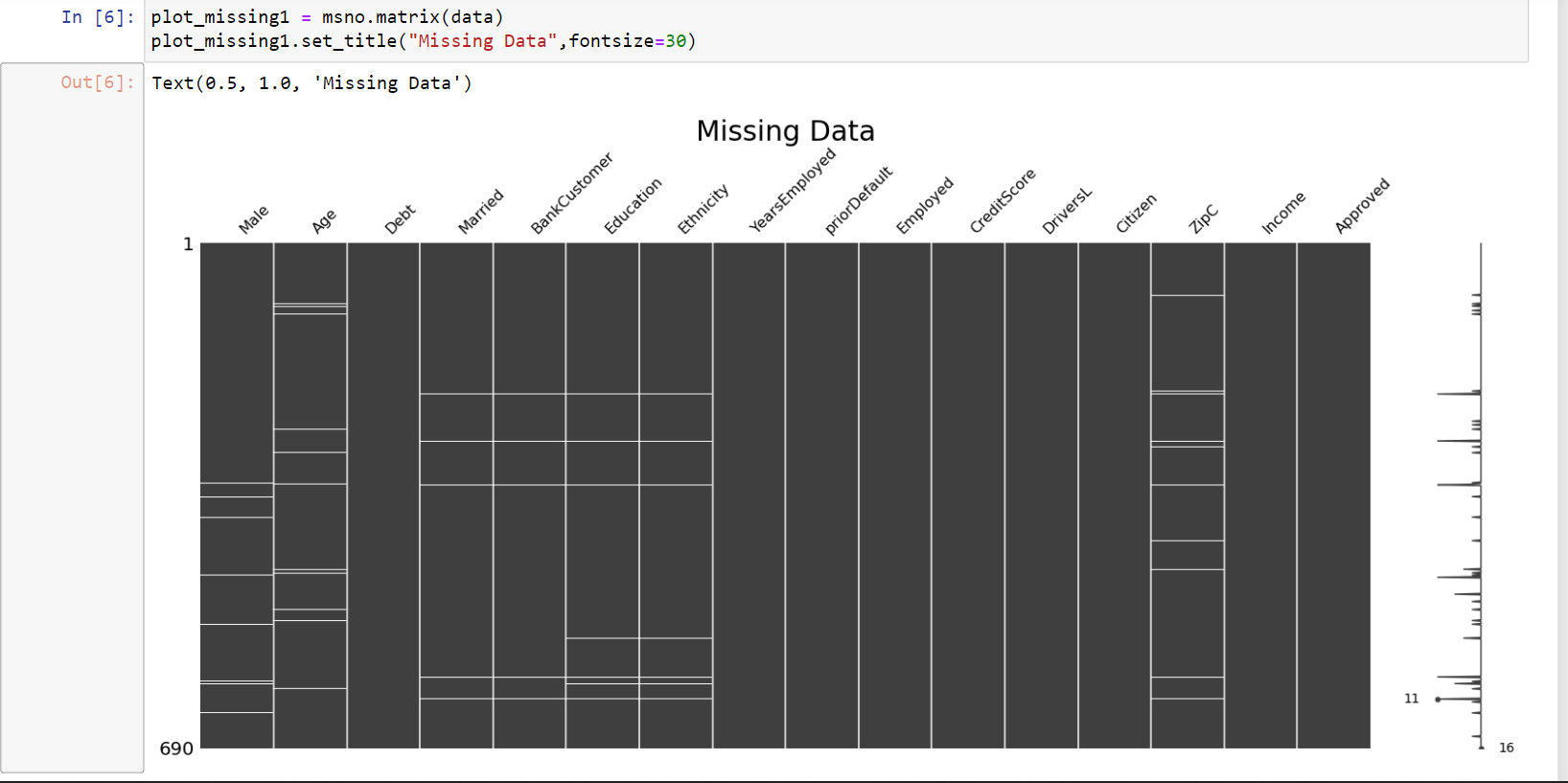
By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model.



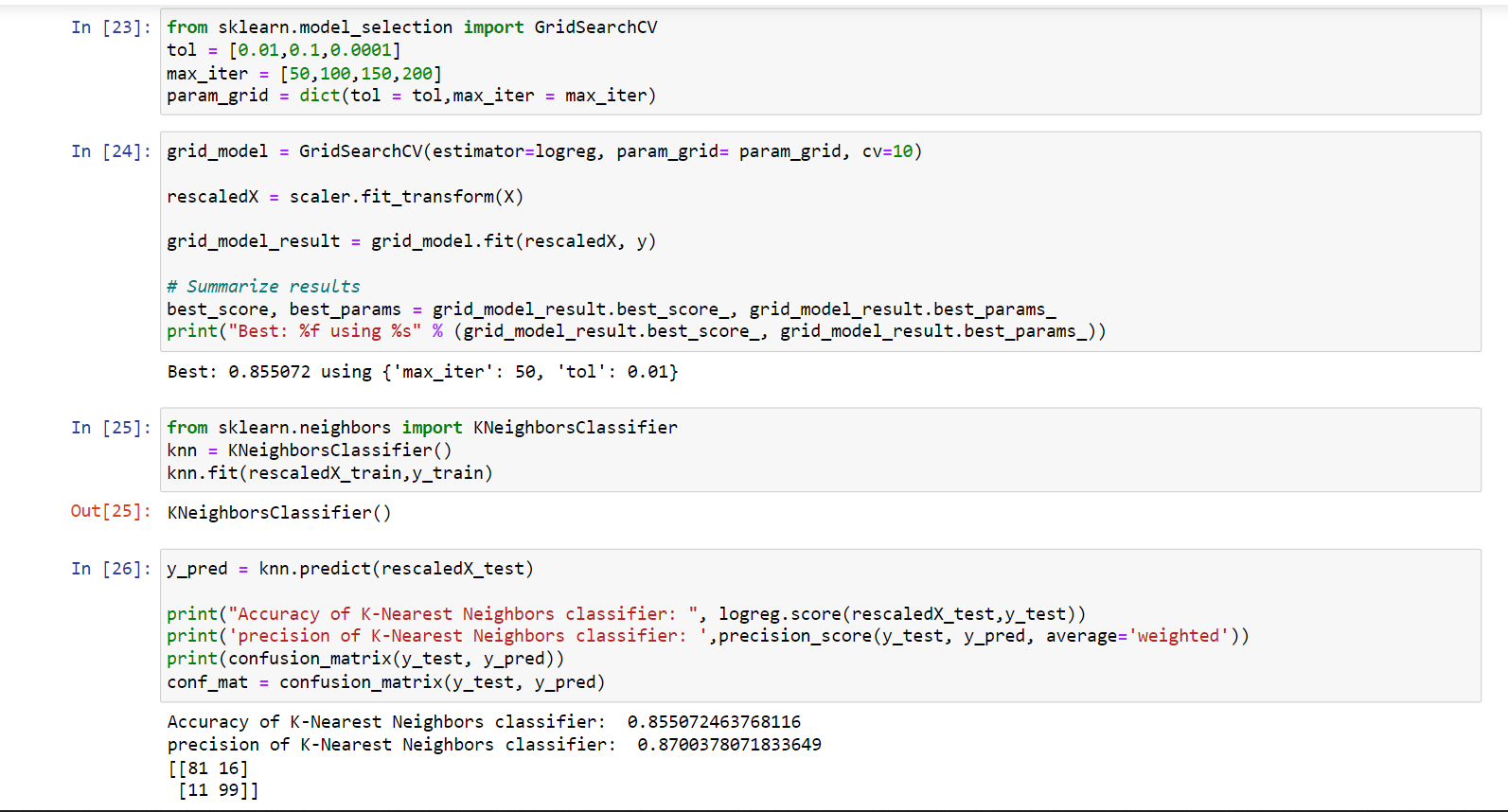
**Code Snippets**

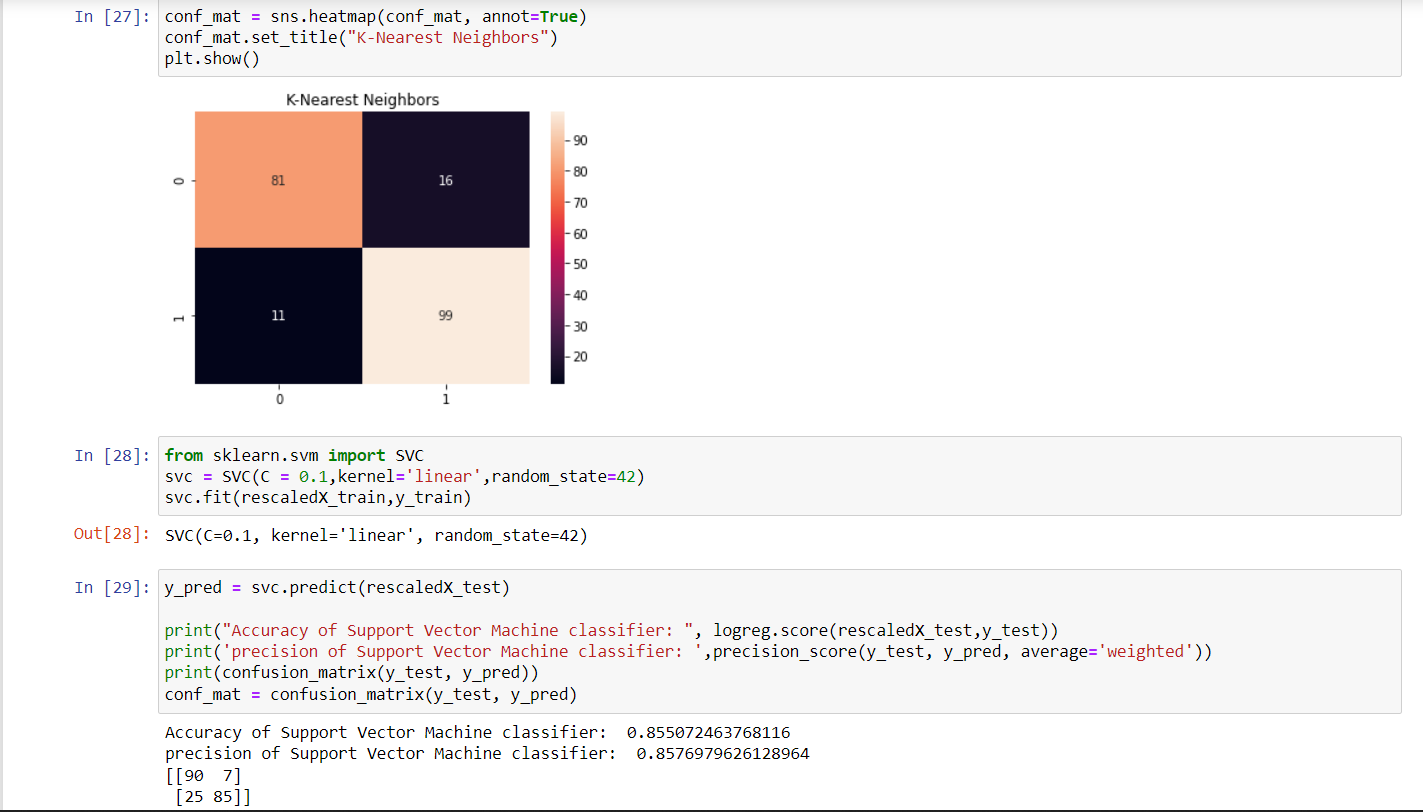




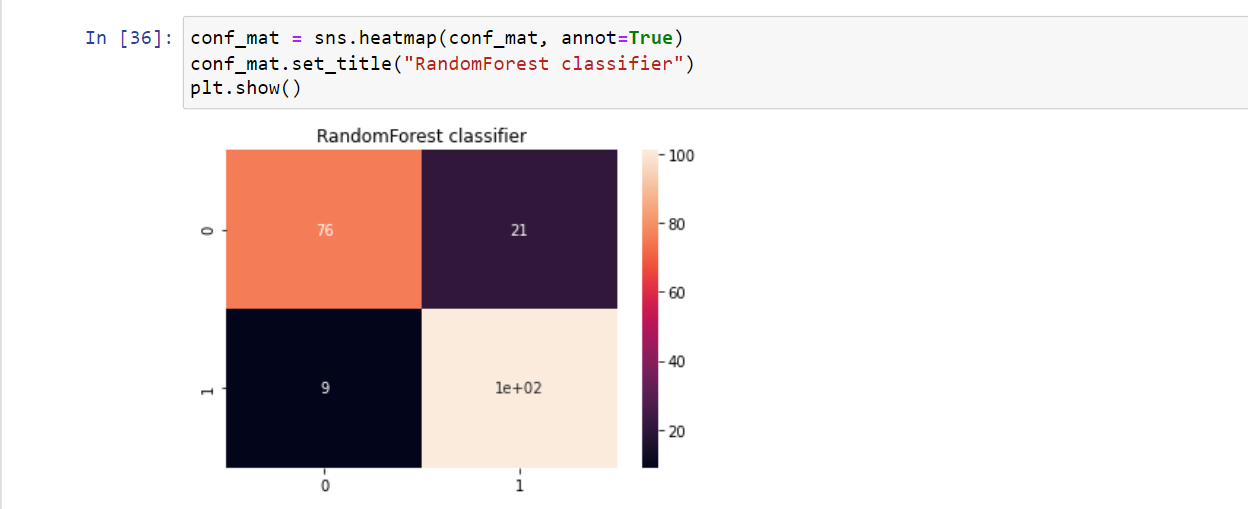
 











**CONCLUSION**

The model we developed, is a predictor and analyser that can define whether a financial institution will issue a card for the consumer. From the observations, we can say the most important keynote features considered by any business host related to the financial background for issuing a card would be prior default, years employed, credit score and debt.

While building this credit card approval predictor model, we tackled some of the most widely-known pre-processing steps such as **scaling**, **label encoding**, and **missing value imputation**. We finished with some machine learning model to predict if a person’s application for a credit card would get approved or not given some information about that person.

Currently, factors considered are regular details related to gender, age of the consumer, his/her credit reports and worthiness, yearly income, and the number of years he/she has been working. Further, to improve this work, various other factors or conditions can be considered like their history related to any offense and their assets which can be both physical and liquid cash. These features can improve the model to be more effective and can help the institutes to make better decisions so that they can avoid experiencing frauds and loss. Various classification algorithms can be used to build a model and compare the rates or levels of accuracy to improve the model for better use.

If you wish to check out more resources related to Machine Learning you can refer to my GITHUB: <https://github.com/shivagowri1928>

**REFERENCES**

UCI MACHINE LEARNING REPOSITORY: <https://archive.ics.uci.edu/ml/datasets/credit+approval>

DATA ANALYSIS PAPER

<http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html>

RESEARCH PAPER

<https://www.researchgate.net/publication/321002603_Credit_Approval_Analysis_using_R>

UNDERSTANDING MACHINE LEARNING ALGORITHMS

<https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>