PROJECT REPORT

Credit Analysis

Foundation of Data Science

Group Number 10

Team Members:

B.Charan Sai	AM.EN.U4CSE19314
R Ashwin	AM.EN.U4CSE19343
Ganeshan M	AM.EN.U4CSE19320
Siva Sai Gopaal	AM.EN.U4CSE19364

TABLE OF CONTENTS

CREDIT ANALYSIS

Abstract

The ability for financial organisations to select potential candidates for a line of credit by identifying the right people with no credit risk is a bit of a challenge. Past demographic and financial data of debtors is essential for building an automated artificial intelligence credit score prediction model based on a machine learning classifier for such a critical choice. Important input predictors (debtor's information) must also be chosen to develop robust and accurate machine learning models. The goal of this computational project is to create a credit scoring prediction model. This study makes use of publicly available credit data.

Introduction

Digital financial services are now one of the most important Big Data sources. In reality, by processing 14 trillion financial transactions every day, worldwide payments income has climbed by 12% in the last two years, reaching 1.9 trillion dollars in 2018 (McKinsey, 2010). The widespread use of financial services has drawn researchers' attention to credit risk management to build models aimed at reducing financial risks while also increasing related revenues.

Lenders' risks in recapturing their investment are associated with credit risk analysis, primarily due to borrowers' failure to repay debt, which is determined by credit risk assessment, also known as *credit scoring*.

Broad Context

Credit analysis comes in handy in the following use-cases

- Credit Status
- Make eligible for loans
- While providing pre approval loans
- Finding any risk involved if score is low
- Knowing Limit of the credit card

Study System

In the study system we analysed various different datasets and took out the best three of them . The parts description and the member in group who focused on that part follows like:

1)Credit Default - B.Charan Sai

2)Credit Eligibility - R Ashwin and Ganeshan M

3)Credit Risk - Sai Gopaal

Default of credit card clients

Analysis of the dataset UCI Default of Credit Card Clients Dataset, which contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 25 variables and 30,000 observations in the dataset:

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX:** Gender (1=male, 2=female)
- **EDUCATION:** (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE:** Marital status (1=married, 2=single, 3=others)
- **AGE:** Age in years
- **PAY_0:** Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
- **PAY_2:** Repayment status in August 2005 (scale same as above)
- **PAY_3:** Repayment status in July 2005 (scale same as above)
- **PAY_4:** Repayment status in June 2005 (scale same as above)
- **PAY_5:** Repayment status in May 2005 (scale same as above)
- **PAY_6:** Repayment status in April 2005 (scale same as above)
- **BILL_AMT1:** Amount of bill statement in September 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May 2005 (NT dollar)

- **BILL_AMT6**: Amount of bill statement in April 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April 2005 (NT dollar)
- **default**.**payment**.**next**.**month**: Default payment (1=yes, 0=no)(Target Variable)

Credit eligibility

Raw Data Contains rows: 251503 columns: 12. The variables are the following:
SeriousDlqin2yrs Person experienced 90 days past due delinquency or worse (Target variable/label)

RevolvingUtilizationOfUnsecuredLines: Total balance on credit cards and personal lines of credit except for real estate and no instalment debt like car loans divided by the sum of credit limits

age: Age of borrower in years

NumberOfTime30-59DaysPastDueNotWorse: Number of times borrower has been 30-59 days past due but no worse in the last 2 years.

DebtRatio: Monthly debt payments, alimony, living costs divided by monthly gross income

MonthlyIncome: Monthly income

NumberOfOpenCreditLinesAndLoans: Number of Open loans (instalment like car loan or mortgage) and Lines of credit (e.g. credit cards)

NumberOfTimes90DaysLate: Number of times borrower has been 90 days or more past due.

NumberRealEstateLoansOrLines: Number of mortgage and real estate loans including home equity lines of credit

NumberOfTime60-89DaysPastDueNotWorse: Number of times borrower has been 60-89 days past due but no worse in the last 2 years.

NumberOfDependents: Number of dependents in a family excluding themselves (spouse, children etc.)

SeriousDlqin2yrs is the target variable (label), it is binary. All of our features are numerical.

The training set contains 150,000 observations of 11 features and 1 label.

Credit Risk

There are 12 variables and 32,581 observations in the dataset

The variables are the following

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_in_*t*rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

- **person_age**: Age in years
- **person_income** : Annual income of the person
- **person_home_ownership** : Home ownership status of the person's current home.
- **person_emp_length** : Length of the person’s employment (Person’s employment history)
- **loan_intent** : Intent of the loan (Reason for taking the loan)
- **loan_grade** : Loan grade (it has various grades like A,B,C,D which are segregated depending on the individual’s professional status)
- **loan_amnt** : Amount of the loan taken by the customer
- **loan_int_rate** : Interest rate of the loan
- **loan_status** : Status of the loan
- **loan_percent_income** : Percent income
- **cb_person_default_on_file** : Historical default
- **cb_person_cred_hist_length** : Length of the credit history

Methods

Data Preprocessing

To improve the performance of the model, we need to do some data processing. Data preprocessing is a necessary step before building a model with these features.The quality of the data should be checked before applying machine learning or data mining algorithms.

Model optimization

Stratified K-Fold Cross-validation

The splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value. This is called stratified cross-validation.
Specifically, we can split a dataset randomly, although in such a way that maintains the same class distribution in each subset. This is called stratification or stratified sampling and the target variable (y), the class, is used to control the sampling process.As the dataset is imbalanced to optimize the results we used Stratified K-Fold.

K-fold Cross-Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models.

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

Credit Default

For the convenience of citizens abroad travelling, credit cards were first issued in Taiwan in 1973. In the year 2000, electronic payment on the internet was established, which led to a rapid expansion of the credit card market. Credit cardholders from different age groups, different education levels, and different gender have different usage behaviours. It is meaningful for banks and financial institutions to investigate the credit card default issue and predict the default of all clients in various conditions.

Model creation

Algorithms used:

1. Logistic Regression
2. Support Vector Machine
3. Stochastic Gradient Descent
4. K- Nearest Neighbour
5. Gaussian Naive Bayes
6. Decision Tree Classification
7. Random Forest Classification

Credit eligibility

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This competition requires participants to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years. This data is from the 2011 Kaggle Competition: *Give Me Some Credit*.

Algorithms used:

1. K-Nearest Neighbours
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Gaussian Naive Bayes

Credit Risk

Credit risk is the possibility of a loss happening due to a borrower's failure to repay a loan or to satisfy contractual obligations. Traditionally, it can show the chances that a lender may not accept the owed principal and interest. This ends up in an interruption of cash flows and improved costs for collection.

Excess cash flows can be written to accommodate additional cover for credit risk. When a lender faces increased credit risk, it can be mitigated through a higher coupon rate, which contributes to more significant cash flows.

Algorithms used:

1. Logistic Regression
2. K- Nearest Neighbour
3. Gaussian Naive Bayes
4. Decision Tree Classification
5. Random Forest Classification

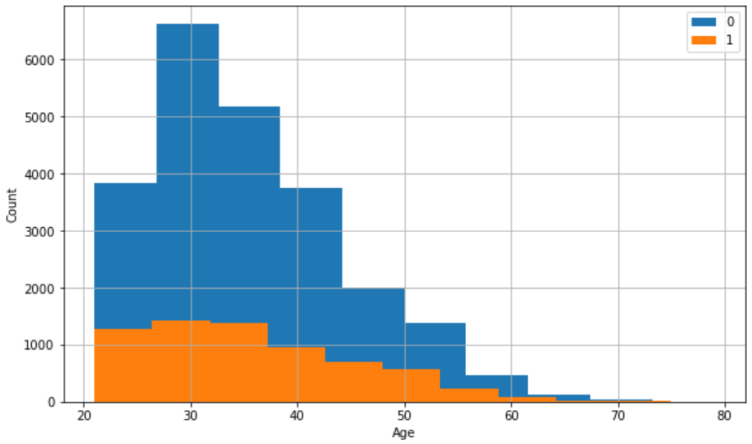
Results

Credit Default Prediction

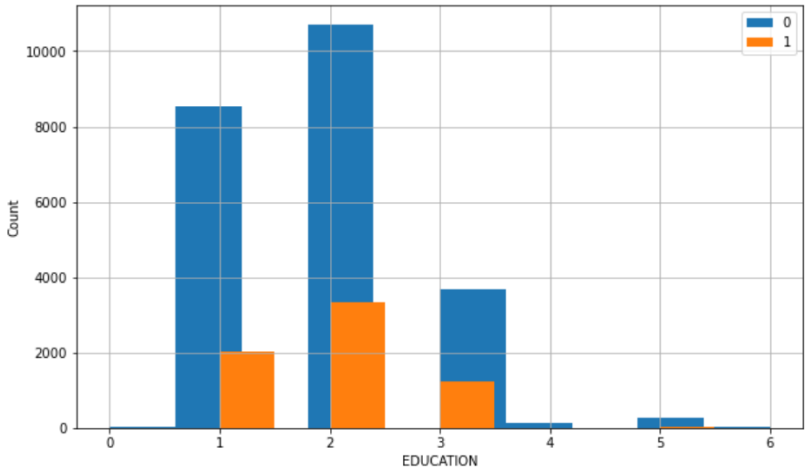
Data Visualization

Bar Graphs between different features and target variable

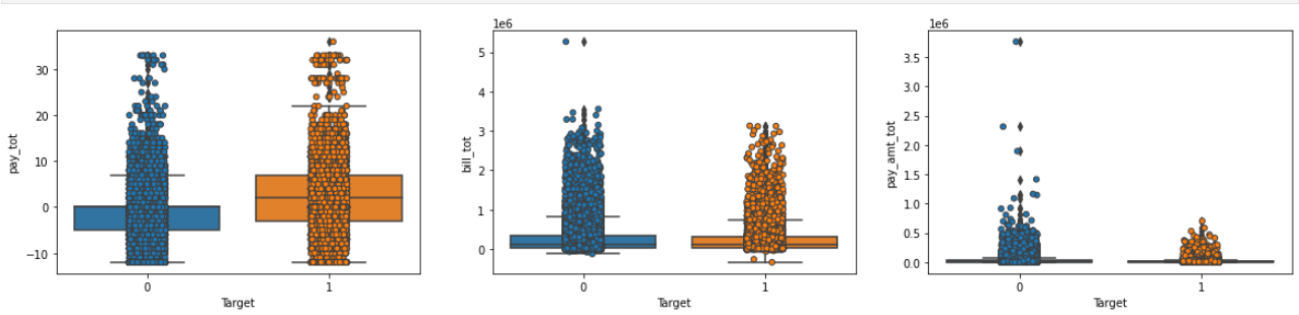
Age:



Education:



Box plot between target feature and pay total,pay amount total,bill amount total respectively.



Performance classifiers :

	Model	Accuracy	Precision	Recall	F1 Score	ROC_AUC
0	Logistic Regression	0.804778	0.685496	0.2245	0.338230	0.597536
1	Support Vector Machine	0.777778	0.000000	0.0000	0.000000	0.500000
2	Stochastic Gradient Descent	0.663111	0.353409	0.6220	0.450725	0.648429
3	K-Nearest Neighbour	0.777778	0.000000	0.0000	0.000000	0.500000
4	Gaussian Naive Bayes	0.765889	0.477568	0.5695	0.519498	0.695750
5	Decision Tree Classifier	0.818111	0.667590	0.3615	0.469024	0.655036
6	Random Forest Classifier	0.816000	0.655797	0.3620	0.466495	0.653857

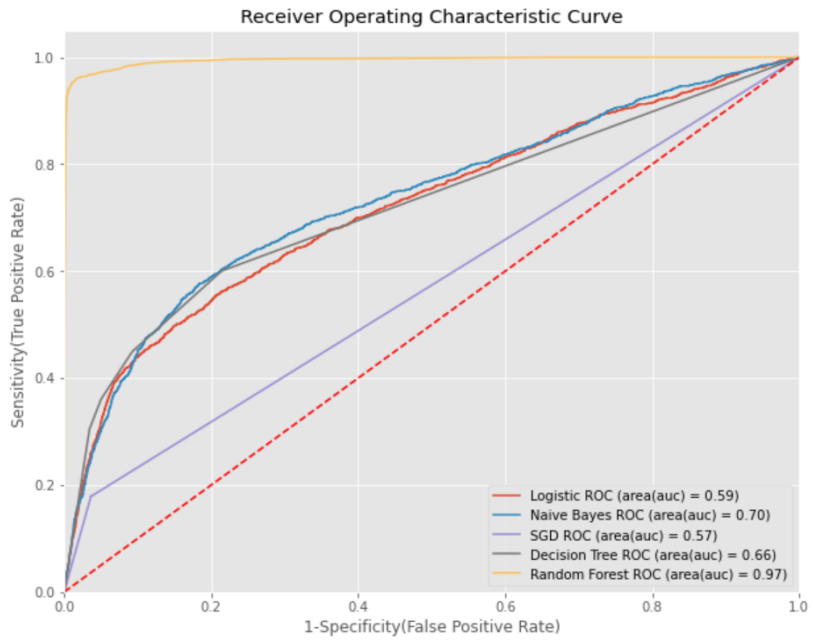
After applying Stratified K-Fold

Optimized performance classifiers:

	Model	Accuracy	Precision	Recall	F1 Score	ROC_AUC
0	Logistic Regression Tuned	0.804000	0.700535	0.197587	0.308235	0.586812
1	Support Vector Machine Tuned	0.779000	0.000000	0.000000	0.000000	0.500000
2	Stochastic Gradient Descent Tuned	0.782667	0.823529	0.021116	0.041176	0.509916
3	K-Nearest Neighbour Tuned	0.779000	0.000000	0.000000	0.000000	0.500000
4	Gaussian Naive Bayes Tuned	0.777667	0.497436	0.585219	0.537769	0.708741
5	Decision Tree Classifier Tuned	0.824667	0.704478	0.355958	0.472946	0.656798
6	Random Forest Classifier Tuned	0.820667	0.661499	0.386124	0.487619	0.665034

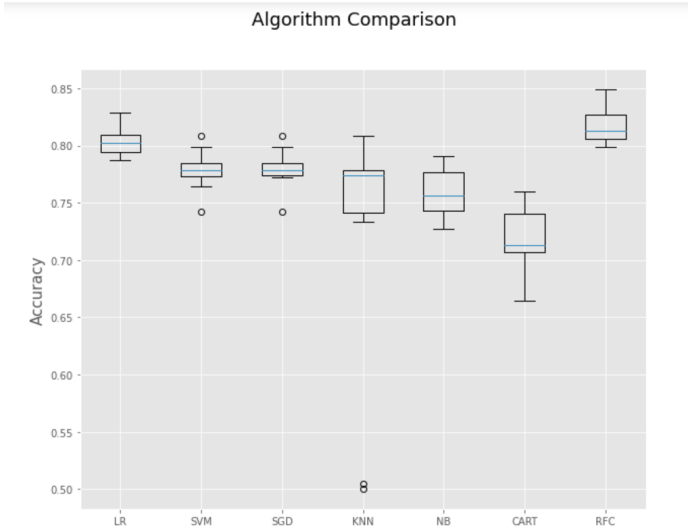
ROC AUC:

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems.The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.



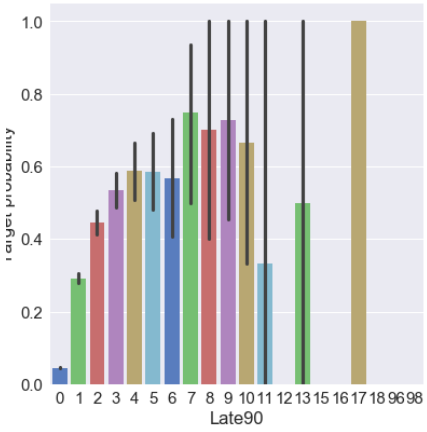
Algorithm Comparison

Box plot accuracy comparison of all algorithms used in credit default prediction.

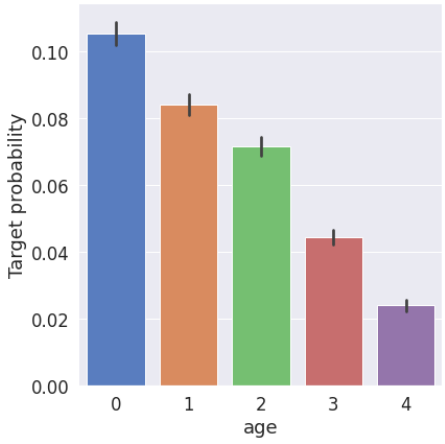


Credit eligibility

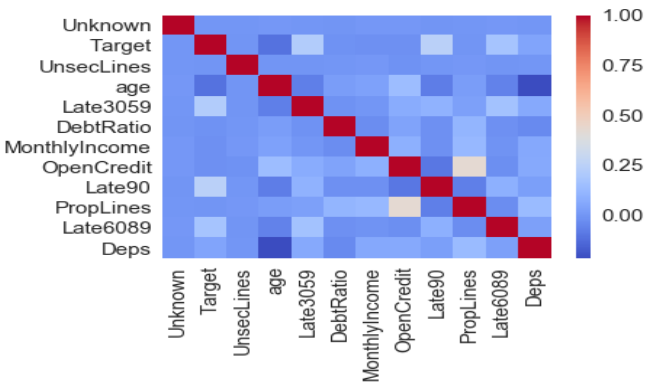
Debt Ratio



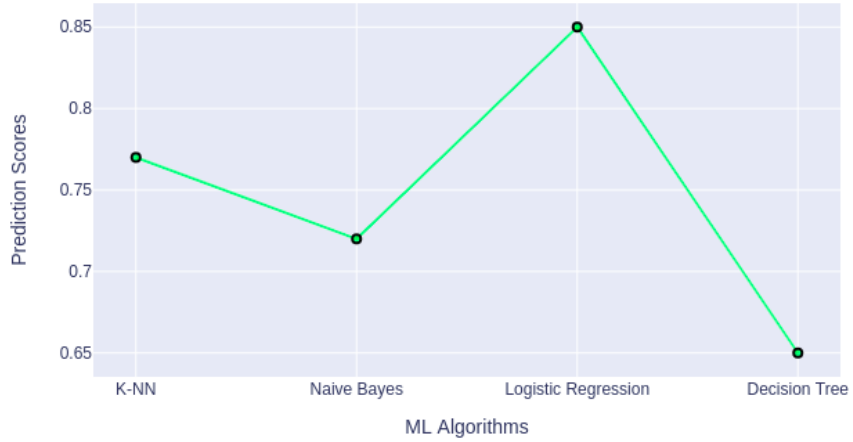
Feature Vs Target



Correlation Matrix

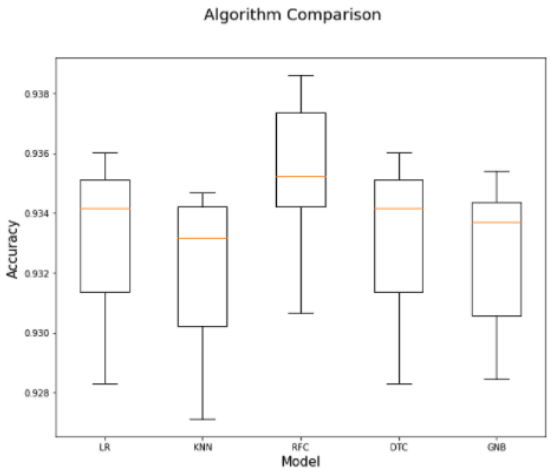


Scatter Plot for Algorithms Comparison



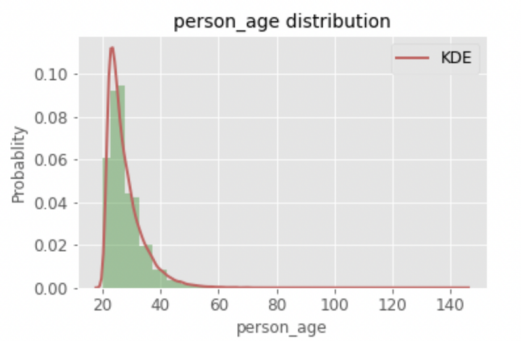
Performance classifiers

Model	Accuracy	Precision	Recall	F1 Score	ROC_AUC
Logistic Regression	0.934134	0.625954	0.040980	0.076923	0.519611
KNN	0.932729	0.491773	0.134433	0.211146	0.562230
Random Forest	0.934502	0.532258	0.181409	0.270593	0.584983
Logistic Regression (Manual)	0.934850	0.055556	0.000777	0.001533	0.499934
Decision Tree	0.897420	0.259276	0.286357	0.272144	0.613818
Gaussian Naive Bayes	0.931892	0.369231	0.023988	0.045049	0.510523

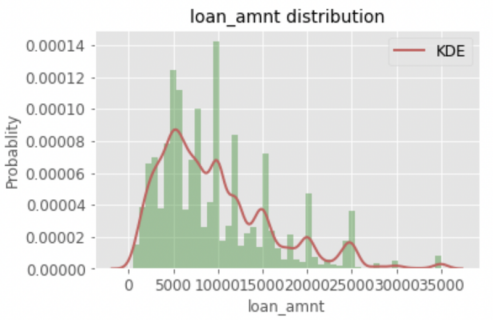


Credit Risk

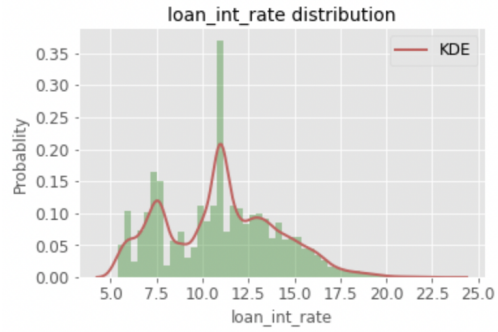
Age



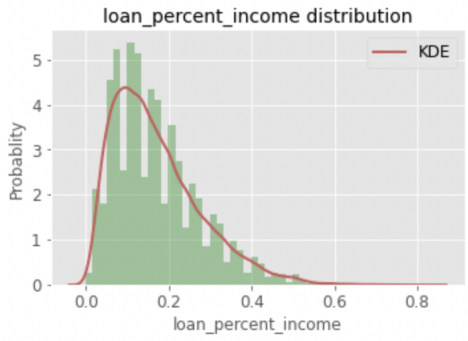
Loan Amount



Interest Rate



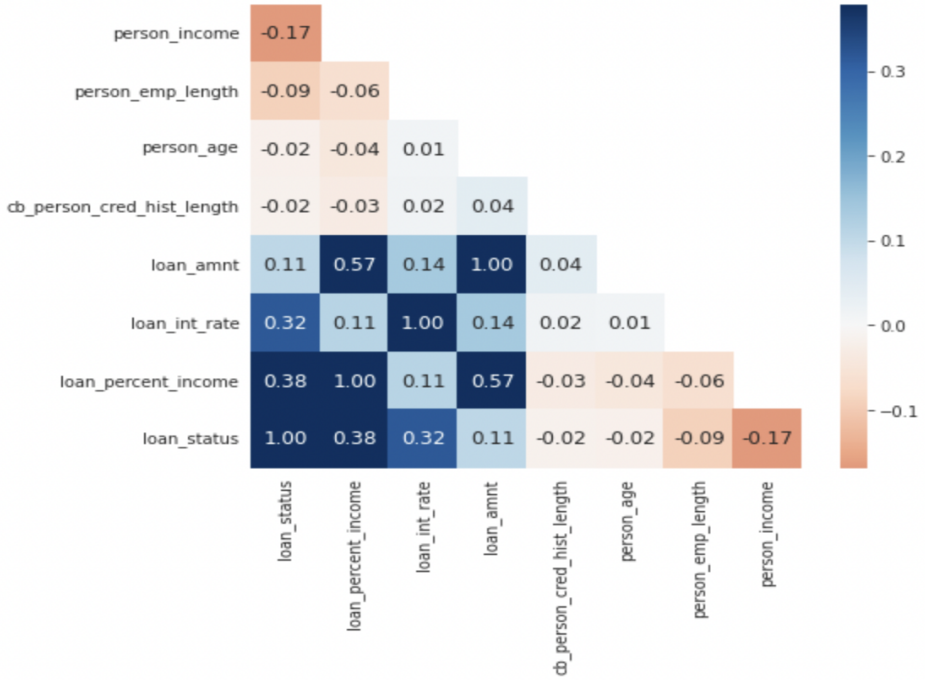
Percent Income



Performance classifiers

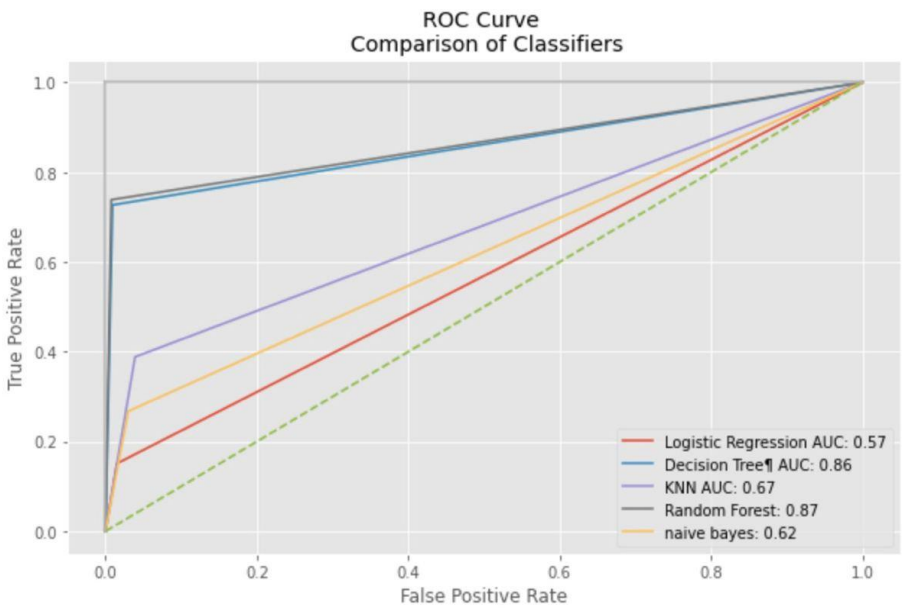
Model	Accuracy	Precision	Recall	F1 Score	ROC_AUC
Logistic Regression	0.802210	0.739130	0.150794	0.250485	0.567927
KNN	0.835465	0.736283	0.388422	0.508557	0.674685
Random Forest	0.999969	1.000000	0.999859	0.999930	0.999930
decision trees	0.932979	0.956976	0.726891	0.826214	0.858859
Logistic Regression Manual	0.771597	0.477636	0.500352	0.488730	0.673822
Naive Bayes	0.817216	0.704071	0.279865	0.400524	0.623519

Correlation Matrix



ROC AUC

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all (various) classification thresholds.



Discussion

Credit Default Prediction

There are 30,000 distinct credit card clients.The average value for the amount of credit card limit is 167,484.The standard deviation is unusually large, max value being 1M.Average age is 35.5 years, with a standard deviation of 9.2.

- All the models accuracy is between 75 and 100%
- We found that for the ROC curve of the Random forest model it consists 0.97 AUC.
- The strongest predictors of default are the PAY_X (i.e. the repayment status in previous months), the LIMIT_BAL & the PAY_AMTX (amount paid in previous months).
- We found that for this data Decision Tree,Logistic Regression and Random Forest Classifier are better.

Credit eligibility

There are 2,51,503 distinct customers whose credit details are available. The average Debt Ratio of customers is 0.3. Average age is 52.3 years, with a standard deviation of 14.7. There are 81 distinct values for the amount of credit limit. Average Number Of Dependents is 0.7.

- All the algorithms have an accuracy between 80% and 100%
- Hence, we found that Random Forest is having the best accuracy.
- Accuracy of the Random Forest is the best when compared to the other 4 algorithms Logistic Regression, KNN, Decision Trees, Naive Bayes.
- Model for Credit risk is Implemented.

Credit Risk

- Details of 32581 customers is available
- Only two columns of data contains null values,
- person_emp_length contains **2.75%** null values and loan_int_rate contains **9.56%** null values
- Most people are 20 to 60 years old
- Most people have less than 40 years of employment
- The cleaned dataset has 32574 rows and 27 columns
- The cleaned dataset has 7 numerical features and 19 categorical features
- Random forest is having the best accuracy
- Maximum ROC is for Random Forest

All the models have an accuracy between 80% and 100%, and the Random forest is having the best accuracy when compared to other models which is 99.99%, which is followed by Decision Tree whose accuracy is 93.29%. Maximum ROC AUC is for random forest which is 0.87.

Literature Cited

https://doi.org/10.1016/j.techsoc.2020.101413

https://doi.org/10.1016/j.eswa.2020.113986

Credit risk Definition | Bankrate.com

What Is Creditworthiness? - Experian

Conclusion

Credit score prediction and risk assessment support financial institutes in defining bank policies and commercial strategies.In this project, we found credit scores, we predicted whether the customer is creditworthy or not, predicted the default of the client and any risk associated or not.This can be used to find Credit Status, Make eligible for loans, While providing pre-approved loans, Find any risk involved if score is low ,Knowing Limit of the credit card.

Project Repositories

Credit Default Prediction - Github
Credit Risk Prediction - Github
Credit Eligibility 1 - GitHub
Credit Eligibility 2 - GitHub