Rainfall in Australia, Used Car Sale Price, Credit Card Default Payment Prediction

Sachin Muttappanavar Student Id: x20144253
MSc in Data Analytics
National College of Ireland Dublin, IRELAND
URL: www.ncirl.ie

***Abstract*—In this paper different Machine Learning methods are applied on Rain in Australia, Used cars, default of credit card clients datasets. Models built using machine learning methods are evaluated using various performance metrics to check their performance and select best model. We followed KDD process to select, process, transfer, and mining data. Machine Learning methods used in this project are Linear Regression, Logistic regression, Decision Tree, Random Forest, Ada Boosting, Gradient Boosting, K Nearest Neighbors. Logistic Regression model is showing good accuracy of 78% for Rain in Australia data set. For Used car dataset, GradientBoost regressor performing good in prediction of reselling price of car with r-squared , 88%. Random Forest model can classify the credit card defaulters accurately with least false negative rate and high positive prediction rate. Important features for solving problem are determined.**

***Index Terms*—Rain in Australia, Used cars price, default od credit card, machine learning, models, regression, classification.**

1. INTRODUCTION

Machine learning is a subbranch of artificial intelligence that learns and improves automatically from historical data. In recent days we can see an application of machine learning in our lives more and more, image recognition, speech recognition, driverless car, and so on. In this project, Machine learning methods like Linear regression, Logistic Regression, decision tree, Random Forest, K Nearest Neighbor are visited to solve real-world problems. Particularly regression and classification problems are solved using different machine learning methods in this work. Models trained using data gathered are compared and evaluated using various evaluation metrics to select the best model. The main goal is to find suitable model that helps in extracting information from given dataset.

A. Rain in Australia

Rainfall forecast is one of the complicated and uncertain tasks which includes a noteworthy effect on human society. Precise and accurate anticipation of rainfall will help humans significantly in preventing and securing from foreseen circumstances also minimize financial losses. In this study, we will be building models with various classification methods of machine learning to predict whether rainfall will occur tomorrow or not using weather information gathered from different cities in Australia on a specific day. Observations were drawn from numerous weather stations. We are using dataset available on kaggle for this work. Dataset has 145460 observations and 23 columns. Research questions for this dataset are:

1. Rain prediction Can we predict next-day rain by using weather data collected from various station across Australia?
2. What are the important features in predicting rain?

B. Used cars dataset.

In the contemporary world, everyone needs a car for various reasons. Some people buy a car to show their social status and some for traveling. The type of car people purchase depends on their choice. The cost of the new vehicle is decided by the carmaker based on various factors like material used in manufacturing, technology, engine specification. Also, a buyer must pay road tax which is set by the Government. In recent days, car prices are increasing as result buyers not able to afford a brand-new car. These developments intercalated interest of buying used cars in customer's mind. Clients will be benefited by purchasing used cars with good condition and features at a cheaper price. Anticipating used car prices is challenging. Many sellers place unrealistic price tag and buyers falls into this trap. Therefore, rises need for a car price anticipating system which systematically predicts prices using vital features influencing final price. In this work we have used various machine learning regression methods to find resale price of car. Models are trained using used cars dataset available on kaggle. Research questions for this dataset are:

1. Can we predict used car prices using historical data of cars?
2. What are the factors that show a significant role in predicting used car prices?

C. Credit card payment defaulters

As per Federal Reserve Economic Data, the percentage of previous due loans within borrowers on credit card loans is at an all-time peak for the last 66 months across all commercial banks and it is expected to rise throughout 2019[2]. This trend will result in huge money losses for commercial banks. Thus, it is very crucial to come up with a risk prediction model which can identify and classify people who have higher chances to default on credit card loans and thereby helping banks in mitigating financial loss. We have trained machine learning models with data uploaded on kaggle. Research questions for this dataset are:

1. How accurately can we predict whether the customer will make payment for next month's credit card bill by building a model using demographic details of the customer and previous payment history?

2. RELATED WORK

Rainfall influences various human activities such as agricultural production, power generation, traveling, forestry, building construction [1]. Also, rain is highly correlated with natural

devastating events like floods, tsunamis, snowslides, landslides. These natural events have adversely affected the world [2]. Therefore, there is an absolute need for an appropriate method to predict rainfall which will help humans in saving themselves from devastating events by taking preventative and relief measures in advance [3]. Recently, a bush fire catch occurred in Australia which damaged nearly 20 million of acres forest land. It killed many human lives and wild animal's lives. This catastrophe has bewildered and led people to research weather prediction patterns. [4]

[5] In this paperwork, three machine learning models are developed using Artificial Neural Network, Support Vector Machines, and Random Forest. Data is collected from autopijaca.ba website through web scrapping which is written PHP. All the models are compared and evaluated to select the best predictive model. The best model has shown an accuracy of 87.38% on test data. The Author Sameerchand Pudaruth [6] worked on the price prediction of used cars in Mauritius in his research paper. He built different models using machine learning algorithms like multiple linear regression, k-nearest neighbors, naive Bayes, and decision trees algorithms. He gathered historical data from daily newspapers. Each model is evaluated and compared to discover the best one. It is concluded in the paper that using the KNN algorithm mean error is significantly reduced than with linear regression. A drawback of this study is small dataset usage.

[7] Author Yuhan Ma in his work on the prediction of credit card default has illustrated five key features that are important in anticipating the default probability of Credit card bills. He used the XGBoost method to solve the problem and has shown model can be used to predict user's readiness to repay credit card debt. He divided data in the ratio of 8:2 as training and testing data. As there was a comparatively small number of variables and samples available for the model, he used 800 trees for XGBoost with maximum depth, learning rate 3, 0.03, respectively. With all these configurations, he can build model with AUC score of 0.779. [8] In this paperwork, it is concluded that the machine learning model built by them can be used by the bank to predict the defaults of customers more accurately. Also mentioned that the same model can be used for preliminary filtering of customers and this would help a bank in minimizing losses in a lending loan. They built a model using the Gradient boost method. It showed an accuracy of 83.7%

3.METHODOLOGY, RESULTS & EVALUATION

In this project we followed Knowledge Discovery in databases methodology to obtain information from dataset and its implementation is as follow:

1. **Data Selection:** This step involves selecting dataset. Focusing on variables and explore relationships between them.
2. **Data pre-process:** Exploratory data analysis techniques are applied to explore variable distribution or frequency or missing values or property.
3. **Transformation:** During this phase data is converted into appropriate form using transformation techniques so that it can be fed to models to understand and learn.
4. **Data Mining:** In this step data mining techniques are used. Models are built and trained with transformed data to discover patterns in the data.
   In this project we have used following machine learning models:
   Linear regression, Penalized models of Linear Regression, Logistic Regression, K Nearest neighbor, Decision Tree, Random Forest, Ada Boosting, Gradient Boosting.
5. **Interpretation/Evaluation:** In this step models performance checked using different metrics. Following metrics are used in this work: Accuracy Area Under Curve (AUC),Recall,Confusion Matrix, Mean absolute error, mean squared error, residual error, r squared error

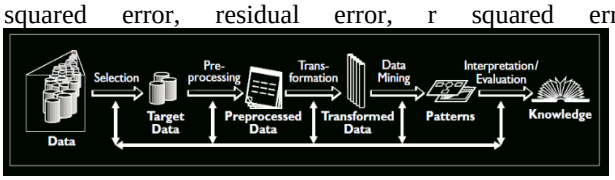


Figure 1. KDD Methodology

A. Rain in Australia

1. Dataset: Rain in Australia dataset contains information about 10 years of daily weather of various places across Australia.

Dimension of dataset:

Number of records: 1,45,461

Number of columns: 23

Table 1. Data set Description

Feature	Description
Date	The date of observation
Location	The common name of the location of the weather station
MinTemp	The minimum temperature in degrees celsius
MaxTemp	The maximum temperature in degrees celsius
Rainfall	The amount of rainfall recorded for the day in mm
Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	The number of hours of bright sunshine in the day.
WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Direction of the wind at 9am
WindDir3pm	Direction of the wind at 3pm
WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Humidity (percent) at 9am
Humidity3pm	Humidity (percent) at 3pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Fraction of sky obscured by cloud at 9am.
Cloud3pm	Fraction of sky obscured by cloud at 3pm.
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
RainToday	1 if precipitation exceeds 1mm, otherwise 0
RISK_MM	The amount of next day rain in mm.
RainTomorrow	The target variable. Did it rain tomorrow?

Figure 2. Columns details

2. Methodology: In this project we followed Knowledge Discovery in databases methodology to obtain information from dataset and its implementation is as follow:

1. *Data Selection:* Data is obtained from kaggle website. It contains information about weather collected from different stations across Australia. Datasets consist of mixture of categorical and numerical features.

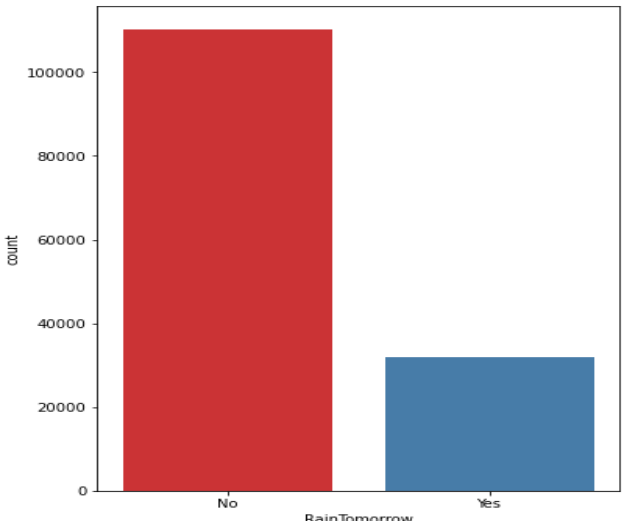


Figure 3 Distribution of classes in RainTomorrow

2. *Pre-Processing:* Dataset consist of missing values.

   a. First investigated summary of dataset. Datasets consist of mixture of categorical and numerical features. We observed there are some missing values. df.describe() command is used to get statistical summary of numerical variables.

   b. Missing values of target variable 'RainTomorrow' are dropped. Frequency distribution of 'RainTomorrow' values is visualized in **Figure 3.** 'Yes' occurs 22.42%, 'No' occurs 77.58% times. We have class imbalance.

   c. Date feature is split into year, month, day and these newly created features are used to other preprocessing work.

   d. Univariate analysis of categorical variable is carried out to check unique and count of values. Missing values of categorical variables are filled by frequently occurring value.

   e. Univariate analysis of numerical variable is carried out to understand the distribution of values and outliers. Outliers are capped by maximum value. As there are outliers in dataset missing values are replaced by median imputation. Median imputation is robust to outliers.

   f. Multivariate Analysis: Heat map of correlation between numerical is plotted to understand interactions between variables in data sets. Input variables with high correlations are removed. Heat map in **Figure 4** and Pair plot in **Figure 5** shows relations between MinTemp, MaxTemp,Rainfall,Evaporation,Sunshine,WindGustSpeed,WindSpeed9am,WindSpeed3pm,Humidity9am,Humidity3pm,Pressure9am,Pressure3pm,Cloud9am,Cloud3pm,Temp9am, Temp3pm variables. Pressure3pm and Temp9am are removed from dataframe as they are highly correlated with Pressure9am and MaxTemp. Presence of multicollinearity causes underestimating independent variable statistical significance.

   g. We observed that there is imbalance in our data set. Imbalance dataset will result in biased model due to less information about minority class. We built models with under sampled and over sampled data. Imblearn library is used to generate samples of minority or majority classes.

3. *Transformation:*

   a. Cardinality of categorical variable is checked. As there was no high cardinality in categorical data, one hot encoding technique is applied to convert string values into numeric.

   b. As pressure9am, Pressure3pm variable consist of bit high magnitude value compared to other, normalization technique is used to bring all numeric values into same scale. Data is then split into train and test with 20% of data as test data.

4. *Data Mining:* We have built different machine learning models of different model family such as Linear classifier, tree-based classifier, ensembles. We used scikit-learn library to build models. Models built with this dataset are – Logistic Regression, Decision Tree, Random Forest, Greadient Boosting with under sampled, over sampled data and data without sampling. Each models are ran through cross validation to check overfitting of model. Hyper parameter tunning technique is applied for decision tree and random forest to detect best parameters for models.

5. *Interpretation/Evaluation:*

Below table shows results of each models.

Under sample: Using imblearner library we under samples are generated to make level par with minority class.

	LogisticRegression	Decision Tree	RandomForest	Gradientdescent
Recall	0.856736	0.842705	0.851208	0.861581
Specificity	0.407686	0.424436	0.446726	0.413411
Precision	0.764402	0.812392	0.817926	0.762315
True Positive Rate	0.856736	0.842705	0.851208	0.861581
False Positive Rate	0.592314	0.575564	0.553274	0.586589
Accuracy	0.718274	0.737016	0.748022	0.720806

In above table, we can see that random Forest model is performing well with high accuracy and reasonable values for specificity and sensitivity. It is selected from above set of models.

Oversample:

For oversampling we have used imblearner library to make minority instances level up with majority class instances. Models built with oversampled data and its results are shown in below table.

	LogisticRegression	Decision Tree	RandomForest	Gradientdescent	Random Forest without tunning
Recall	0.857756	0.789276	0.803932	0.861581	0.849534
Specificity	0.400264	0.348349	0.606838	0.413411	0.448534
Precision	0.752472	0.914043	0.964529	0.762315	0.822598
True Positive Rate	0.857756	0.789276	0.803932	0.861581	0.849534
False Positive Rate	0.599736	0.651651	0.393162	0.586589	0.551466
Accuracy	0.711382	0.744189	0.790147	0.720806	0.749534

From above set of models, we have picked random forest model as it is having highest accuracy and fair values of recall and specificity.

Adjusting threshold:

	0.5 threshold	0.6 threshold	0.7 threshold	0.8 threshold
accuracy	0.711382	0.757340	0.771124	0.781673
recall	0.857756	0.843128	0.826558	0.793050
precision	0.752472	0.844008	0.891908	0.972013
specificity	0.400264	0.460126	0.487527	0.564880
roc_auc_score	0.661079	0.661079	0.661079	0.661079

Specificity and recall are inversely proportional as one increase other value will go down. We can observe in above table specificity is increasing as threshold increases whereas recall value is decreasing. We are setting threshold to 0.8 for which recall, and specificity are high. Recall explains what percentage of prediction that are correctly identified it will rain tomorrow. Conversely, specificity explains percentage of predictions that says it will not rain tomorrow. It would be good for people to be notified if there is any chance of rain occurrence, that is why we are selecting threshold of 0.8 which is showing reasonable recall and specificity value.

```
array([[21429,   617],
       [ 5592,   801]], dtype=int64)
```

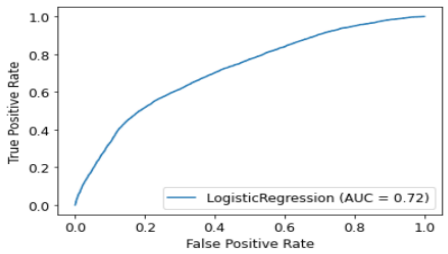


Figure 6. ROC curve

Above **Figure6** is roc_auc curve plot which helps to find threshold level at which specificity and sensitivity balances. For perfect classifier Area under curve should be equal to 1. Higher the value of AUC better is model in predicting rain tomorrow. In this case we have fair AUC value of 0.72.

Comparing models from each section:

We are rejecting model built with under sampled data as considerable amount of data is deleted from dataset which may be case of loss of important data.

While random Forest model of under sample is built with tunned parameters using RandomSearchCV. When we looked at accuracy on train and test set for this model, model is overfitting train data as it is showing 100 % accuracy for train and 80 % on test, thereby not able to classify target value with new data. That is why rejecting this model.

```
Accuracy on test data: 79.01473328879356
Accuracy on train data: 100.0
```

Logistic Regression model built with adjusted threshold of 0.8 is chosen as final model which is giving accuracy rate of 78% .

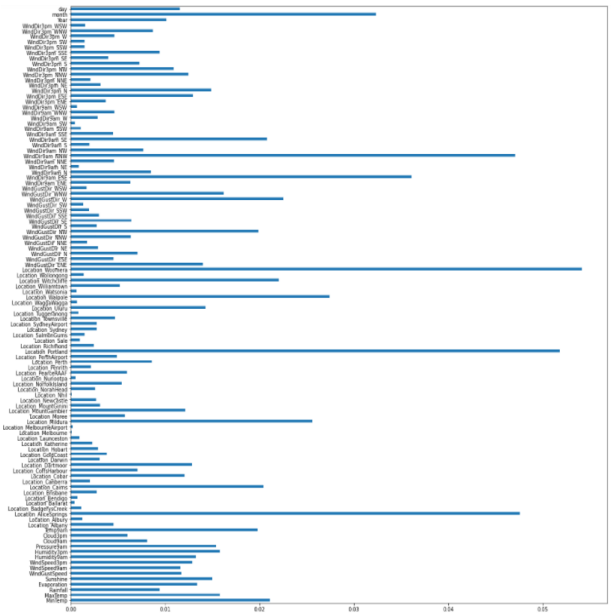


Figure 7. Feature importance

B. Car sale price:

1. Dataset: Data available in this dataset is gathered from world's biggest used car sale collection 'Craigslist'. It contains information that craigslist website showcases on car sale. For this work 14 features are utilized and same are mentioned below.

Dimension of dataset:

Number of records: 4,58,261

Number of columns: 14(For this work we will be using 14 features out of 26.)

Region – Branch of CraigsList company in US.
Year – Year car brought.
Manufacturer – Vehicle manufacturer.
Condition – Current condition of car.
Cylinders – Number of cylinders used in car.
Fuel – Fuel type of car.
Odometer – Total run by car.
Title Status - legitimate document that specifies the person or business that legally owns the car.
Transmission – Tells about gear type(Automatic or manual, other).
Drive – Type of drivetrain (rwd – rear wheel drive, 4wd – four wheel drive, fwd – front wheel drive).
Type of car (example- Sedan, SUV).
Paint color – Car color.
State – US state car belongs to.
Price – Final price of car (Target).

Figure 8. columns details of Car sale dataset

2. Methodology: We followed KDD process to discover insight of dataset.

1. *Data Selection:* This dataset is available on kaggle. Dataset comprised of categorical and numerical data.
2. *Pre-Processing:*
   a. Dropped unnecessary columns like 'id','url','region_url','VIN', 'image_url', 'description','lat','long','region','Unnamed: 0', 'posting_date' from dataset.

b. Outliers from price column are removed. Shape of dataframe post removal of outliers: rows 337620, columns 15.

c. Outliers of 'odometer' column are removed. Dataset is filtered to keep rows with odometer less than 3000000.

d. As there were less observations with year value less than 1940, dataset is filtered to have year value greater than 1940.

e. Null values of condition column are replaced based on odometer reading.

f. Dropped rows with null values for columns:'title_status','fuel','transmission','model','manufacturer'.

g. Null values of 'paint_color', 'drive', 'type', 'cylinders' are replaced by forward fill.

h. Pairplot in **Figure 8** shows the linear relation between year, price, odometer after processing.

i. Then bar plot, box plot, strip plot is visualized to check relation of categorical variable with respect to target variable in **Figures 9 to 13**.

j. Histogram of price column distribution is shown in **Figure 14.**

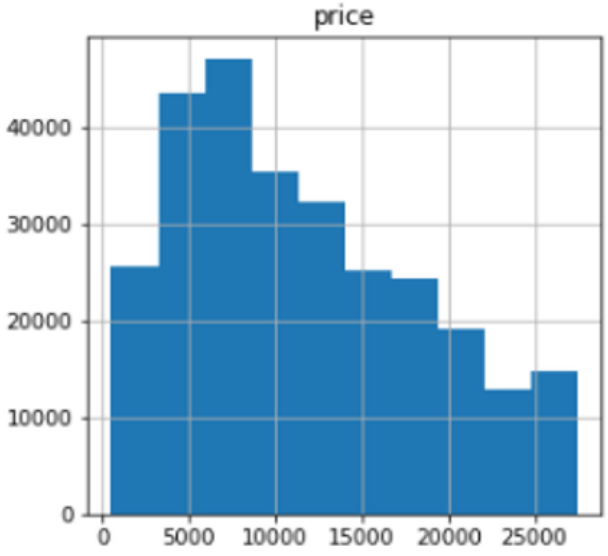


Figure 14. Distribution of Price

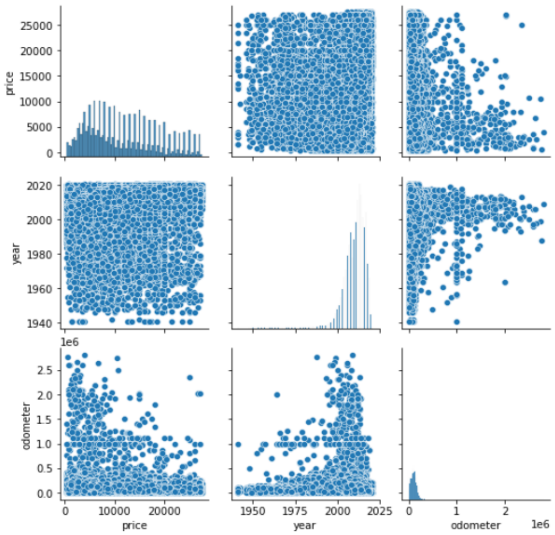


Figure 15. Linear relationship between variables

3.Transformation: Categorical variables are converted into integer values using LabelEncoder of Preprocessing module of sklearn.

4.Data Mining: For this dataset we built 7 different models and they are RandomForestRandomForest,LinearRegression,Ridge Regression,Lasso Regression ElasticNet,Gradient Boosting,Ada Boosting. Data is split into train and test with test size 20 %. These models are trained with trained data and tested with test data to check how model is behaving with respect to out of sample data.

5.Inerpretation and Evaluation:

	RandomForest	LinearRegression	Ridge Regression	Lasso Regression	ElasticNet	Gradient Boosting	Ada Boosting
Mean Absolute Error	1.691080e+03	4.259970e+03	4.259970e+03	4.260190e+03	4.385740e+03	1.807220e+03	4.259970e+03
Mean Squared Error	7.802950e+06	3.017796e+07	3.017796e+07	3.017825e+07	3.113619e+07	7.660862e+06	2.673294e+07
Root Mean squared error	2.793380e+03	5.493450e+03	5.493450e+03	5.493470e+03	5.579980e+03	2.767830e+03	5.170390e+03
R squared error	8.397835e+01	3.803617e+01	3.803617e+01	3.803557e+01	3.606866e+01	8.427010e+01	4.510977e+01

As we can see in above table RandomForest and Gradient Boost regressor giving good results with less Root mean squared error, mean squared error, and Mean absolute error. Even R squared value is around 84 for both models.

We ran cross validation with 10 splits for both models. Both models are producing mean score of 0.83 and 0.84 with standard deviation 0.007 and 0.005, respectively. We have chosen gradient boosting as final model as it is showing good R squared value and making less error in making prediction. **Figure 17** shows distribution plot of difference between actual and predicted values. We can see that there is not much difference between actual and predicted values as results are concentrated around zero mean.

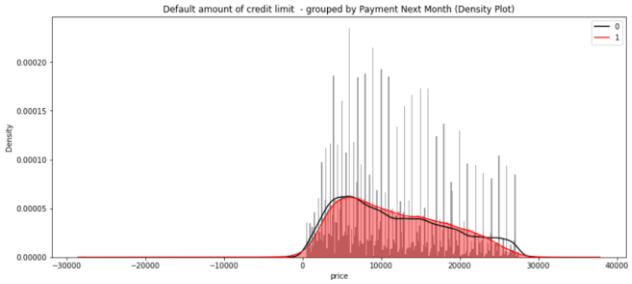


Figure 16. Distribution of actual and predicted values

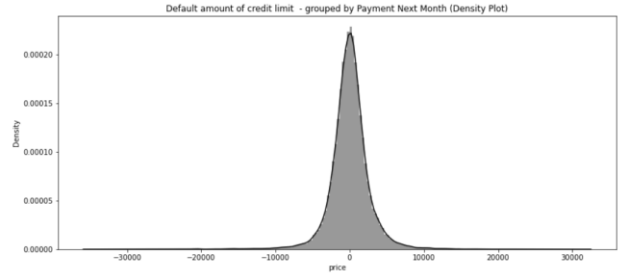


Figure 17. Distribution of difference between actual and predicted values

	Mean	Standard Deviation
Train	0.839588	0.005553
Test	0.784049	0.008238

We can also see that there is not much difference between the accuracy on train and test for gradient boosting model. That means, model is not overfitting to train data and able to estimate population parameter of target value for given out of sample data.

Important features in predicting target value are visualized in below **Figure 18**.

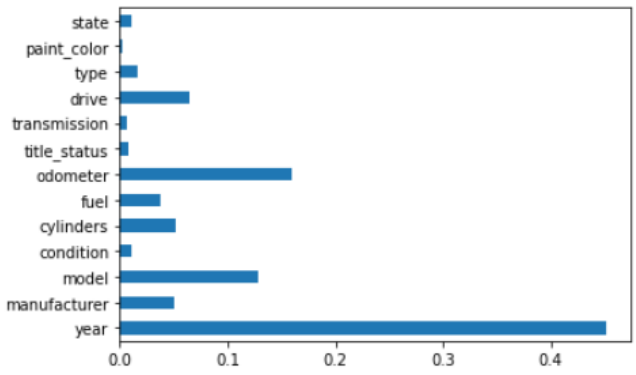


Figure 18. Feature importance

C. Default of credit card:

1. Dataset: This dataset consists of information about default payments, customer information, payment history, bills, credit data of credit card holders in Taiwan from April 2005 to September 2005. There are 25 variables in the dataset and are shown in below **Figure 19**.

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 of client in years
PAY_0: Status of repayment 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: Status of repayment in August, 2005 (scale same as above)
PAY_3: Status of repayment in July, 2005 (scale same as above)
PAY_4: Status of repayment in June, 2005 (scale same as above)
PAY_5: Status of repayment in May, 2005 (scale same as above)
PAY_6: Status of repayment in April, 2005 (scale same as above)
BILL_AMT1: Bill statement in September, 2005 (NT dollar)
BILL_AMT2: Bill statement in August, 2005 (NT dollar)
BILL_AMT3: Bill statement in July, 2005 (NT dollar)
BILL_AMT4: Bill statement in June, 2005 (NT dollar)
BILL_AMT5: Bill statement in May, 2005 (NT dollar)
BILL_AMT6: Bill statement in April, 2005 (NT dollar)
PAY_AMT1: Payment of previous bill in September, 2005 (NT dollar)
PAY_AMT2: Payment of previous bill in August, 2005 (NT dollar)
PAY_AMT3: Payment of previous bill in July, 2005 (NT dollar)
PAY_AMT4: Payment of previous bill June, 2005 (NT dollar)
PAY_AMT5: Payment of previous bill in May, 2005 (NT dollar)
PAY_AMT6: Payment of previous bill in April, 2005 (NT dollar)
default.payment.next.month: Default payment (1= default payment, 0=payment will be made)(Target)

Figure 19. details of columns in dataset3

Dimension of data:

Number of records: 30,001

Number of columns: 25

2. Methodology: We followed KDD methodology to fetch information from this dataset.

1. *Data Selection:* This dataset is picked from kaggle. All variables in this dataset are in numerical.
2. *Pre-Processing:*
   a. Dropped ID column as it does not explain anything about target variable.
   b. There are no null or missing values in the entire dataset.
   c. Statistical summary of dataset is as follow:
      - Mean value of credit card balance is around 167000. Standard deviation is very large and 1M is highest value.
      - Mean Education level is 1.85 which means credit card is mostly used by people whose highest education is graduate school or university.
      - Larger portion of customers are either single or married.
      - Mean age is around 35.5 with standard deviation 9.2.
   d. We examined distribution of credit limit in **Figure 20.**
   e. Checked input variables with zero variance using VarianceThrshold function of sklearn library.
   f. Then checked distribution of target variable classes in **Figure 21**. We can see that target variable classes distribution is not highly imbalanced. Data is not highly unbalanced with respect to independent variable default.payment.next.month.
   g. In **Figure 22** we have visualized how credit balance is with respect to sex. Both male and

female are having almost same credit balance. Female is having highest outliers (1 million). Mean of male is lower than mean of female. Male is having larger third quartile and smaller first quartile value compared to female.

h. **Figure 23** and **Figure 24** are histograms each input variables to check distribution of values and heatmap to show correlation between variables.

3.Transformation: Scaled data to bring input variables into on same scale using StandardScaler function of sklearn library. If feature scaling is not applied, then variables with higher magnitude will dominate in calculating distance therefore resulting in wrong prediction.

We applied Principal component analysis technique applied to decrease dimensionality of big dataset. **Figure 25** is scree plot to select number of components for PCA. Its plot with factors on x-axis and variance on y-axis. We must consider the factors above level off point as optimal number of components. We transformed our data into 4 components. Used transformed data for model building.

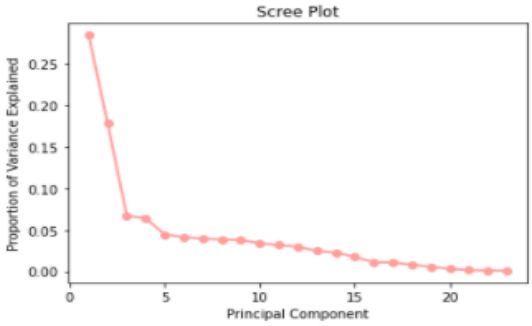


Figure 25. Scree plot

4.Data Mining: We built models with and without PCA transformed data. In total 7 models of distance, tree and regression-based algorithms. Data is split into train and test dataset with 20 % test size. Model is first given with train data to learn pattern and later tested using holdouts. Reason for data split is avoid overfitting of model. Make model generalized. Each models' results are interpreted in next step.

5.Interpretation and Evaluation:

Without PCA:

	LogisticRegression	RandomForest	Ada boosting	Decision Tree
Recall	0.818978	0.840422	0.834948	0.831351
Sensitivity	0.679217	0.668610	0.665612	0.644086
Precision	0.969744	0.951562	0.954972	0.952983
True Positive Rate	0.818978	0.840422	0.834948	0.831351
False Positive Rate	0.320783	0.331390	0.334388	0.355914
Accuracy	0.808667	0.820778	0.817111	0.812000

Positive predictive rate is high for Logistic regression model, but its recall value is lesser than Random Forest. It would be good to have high recall value to alert bank if there is any doubt about defaulter. So, selecting Random Forest model from above set as final model.

With PCA: We have built 3 models with PCA transformed data. Results of each model is shown below table.

	LogisticRegression	RandomForest	KNN
Recall	0.803179	0.823993	0.828928
Specificity	0.621622	0.602724	0.600000
Precision	0.976136	0.950284	0.944318
True Positive Rate	0.803179	0.823993	0.828928
False Positive Rate	0.378378	0.397276	0.400000
Accuracy	0.794222	0.802333	0.804000

Random Forest classifier parameters are tunned using hyperparameter technique. Model developed using best parameters as below:

```
RandomForestClassifier(max_depth=10, max_features='sqrt', min_samples_leaf=2,
                       min_samples_split=5, n_estimators=1000)
```

For KNN, checked accuracy with different neighbor value. Selected model with good accuracy. **Figure 27** shows accuracy of KNN model with varying neighbor value.

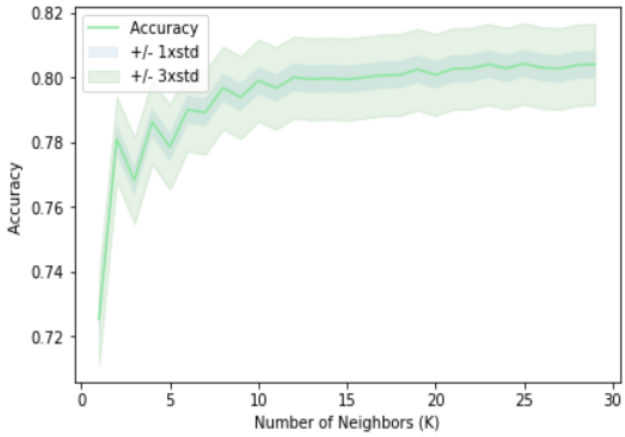


Figure 27. No. of Neighbors vs Accuracy

We can see that after dimensionality reduction, we are still get almost same results as without dimensionality reduction for each model. Here also, Random Forest model wins race as it has good results than other models.

Adjusting threshold:

Logistic Regression threshold value to decide defaulter or not is varied to tune the recall, precision, specificity values. In the starting, we have used default threshold of 0.5. If observe results, recall is high which indicates ability to find defaulter is high whereas specificity is low that means low ability in predicting credit payments. It is not good to have less percentage in predicting non defaulters. So, we tried to bring the specificity value bit higher by increasing threshold. As we can see in below table as threshold value increases, specificity is improving. Model is giving good result with 0.7 as threshold. Its good to have bit higher recall than specificity as it would help bank to examine if there is slight doubt about defaulter. So, using logistic model with 0.7 threshold from below set.

	0.5 threshold	0.6 threshold	0.7 threshold
accuracy	0.631333	0.788333	0.800000
recall	0.861640	0.843570	0.814980
precision	0.629830	0.895455	0.962926
specificity	0.323819	0.518009	0.617302
roc_auc_score	0.633282	0.633282	0.633282

Comparing all three final models from above sections:

Logistic regression model threshold value is tunned to get good results for precision and specificity, but area under curve is low as 0.59. However, Random Forest with tunned parameter is showing good results with reasonable area under curve (0.75) in True Positive rate and False Positive rate plot (**Figure 29**). This model is trained using PCA transformed data thereby addressing problem of curse of dimensionality.

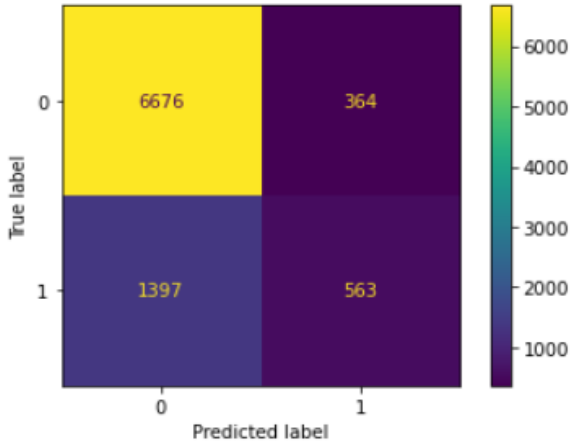


Figure 28. Confusion Matrix

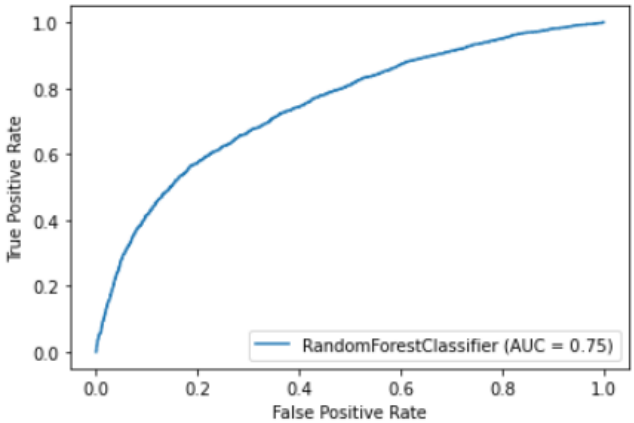
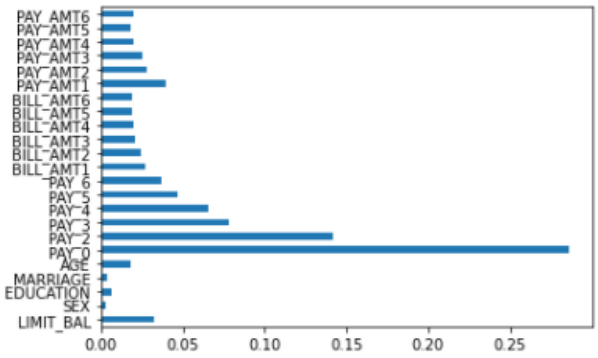


Figure 29. ROC curve

Feature importance is shown **in Figure 30** below.



4 CONCLUSION AND FUTURE WORK

To conclude, we have visited different machine learning methods belonging to various algorithm family. We can use them to predict target variable in regression as well as in classification problems with high accuracy. We also applied transformation, feature selection, hyper parameter tunning techniques to improve the models. Each model is evaluated for their performance using multiple metrics.

Dataset1: Rain in Australia: For this model, Logistic regression model is predicting rain tomorrow correctly with 78% of accuracy. For this dataset we would suggest collecting data with balanced classes in the target variable and use dimensionality reduction techniques to reduce curse of dimensionality for future work.

Dataset2: Car's sale price: For this dataset, Gradient Boosting model can predict resale price of cars more accurately with 88 % accuracy. For improving accuracy of model, we would suggest applying hyper parameter tunning technique to find best parameters for models.

Dataset3: Credit card defaulter: Random Forest algorithm can classify given input data as payer or defaulter with accuracy of 80 %, recall of 82 %, specificity of 60 %. For this dataset we have gathered 30000 records of data, we could have achieved more accuracy value if trained data with some more data. For future work we would recommend increasing dataset size.

REFERENCES

[1] World Health Organization: Climate Change and Human Health: Risks and Responses. World Health Organization, January 2003.

[2] Alcntara-Ayala, I.: Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. Geomorphology 47(24), 107124 (2002).

[3] Nicholls, N.: Atmospheric and climatic hazards: Improved monitoring and prediction for disaster mitigation. Natural Hazards 23(23), 137155 (2001).

[4] P. Lynch, "The origins of computer weather prediction and climate modeling," Journal of Computational Physics, vol. 227, no. 7, pp. 3431– 3444, 2008.

[5] Car Price Prediction using Machine Learning Techniques Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric International Burch University, Sarajevo, Bosnia and Herzegovina.

[6] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol

4.7 (2014): 753-764.
[7] Yuhan Ma - Prediction of Default Probability of Credit-Card Bills. University of Wisconsin-Madison, Madison, WI, USA.
[8] Merikoski, M., Viitala, A. and Shafik, N. (2018) Predicting and Preventing Credit Card Default.
[9] Practical Statistic for Data Scientist by orielly book.
[10] Feature engineering for machine learning by orielly book.
[11]. Introduction to statistical Learning with application in R.
[12] Rain Prediction Data set
[13] Used car dataset
[14] Credit card client's datase
[15] B. P. Salmon, W. Kleynhans, C. P. Schwegmann and J. C. Olivier, "Proper comparison among methods using a confusion matrix," 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 3057-3060, doi: 10.1109/IGARSS.2015.7326461.
[16] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.
[17] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum," 2009 International Joint Conference on Artificial Intelligence, 2009, pp. 353-356, doi: 10.1109/JCAI.2009.146.
[18] R. K. Grace and B. Suganya, "Machine Learning based Rainfall Prediction," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 227-229, doi: 10.1109/ICACCS48705.2020.9074233.
[19] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.
[20] S. K. Mohapatra, A. Upadhyay and C. Gola, "Rainfall prediction based on 100 years of meteorological data," 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 162-166, doi: 10.1109/IC3TSN.2017.8284469.

APPENDIX

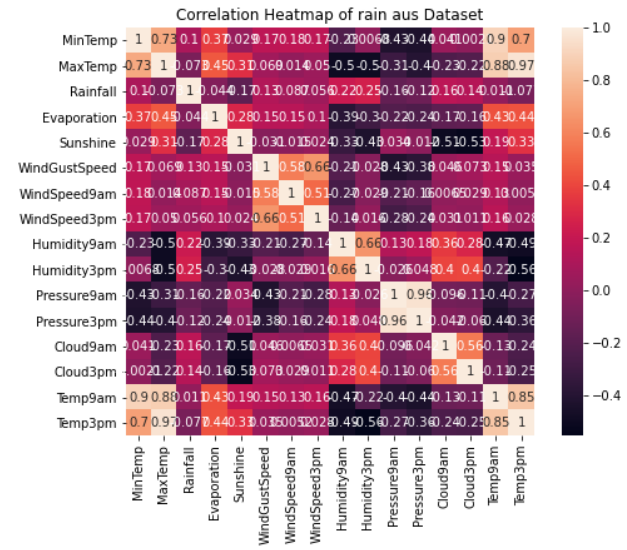


Figure 4. correlation map between variables

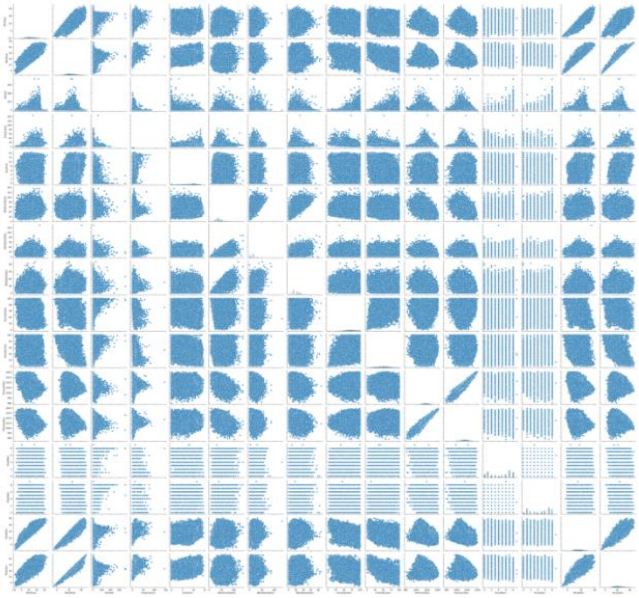


Figure 5. linear relatin between variables

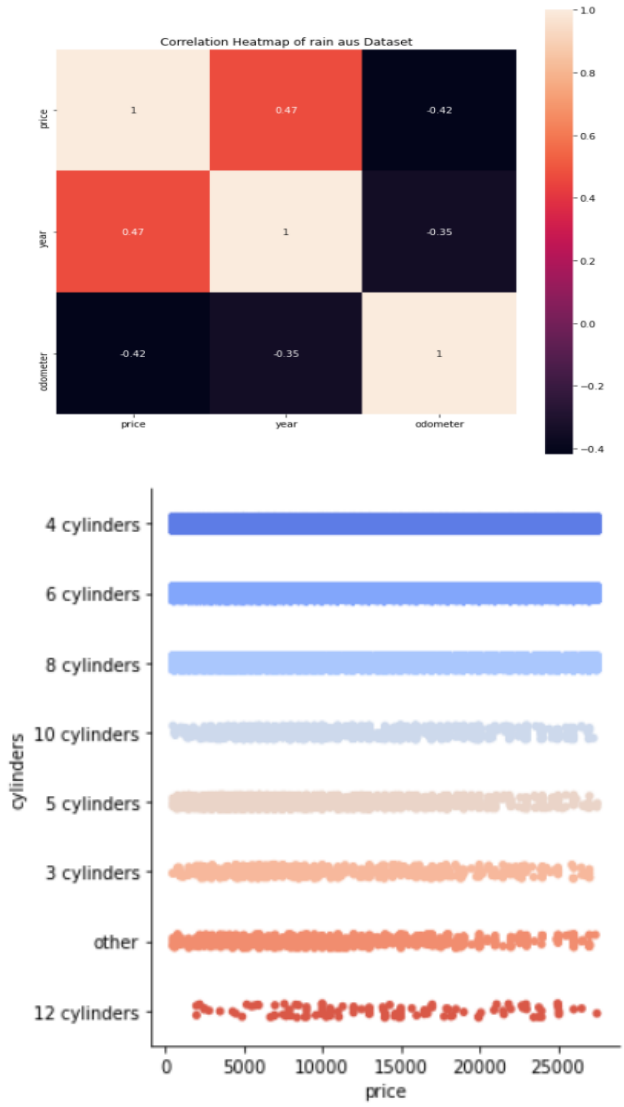


Figure 9. Cylinders vs price

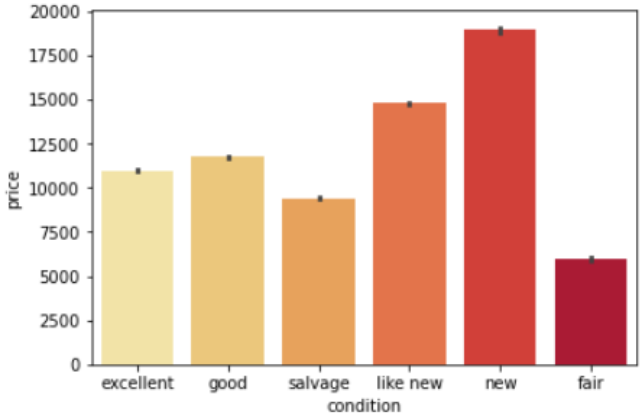


Figure 10. Condition vs price

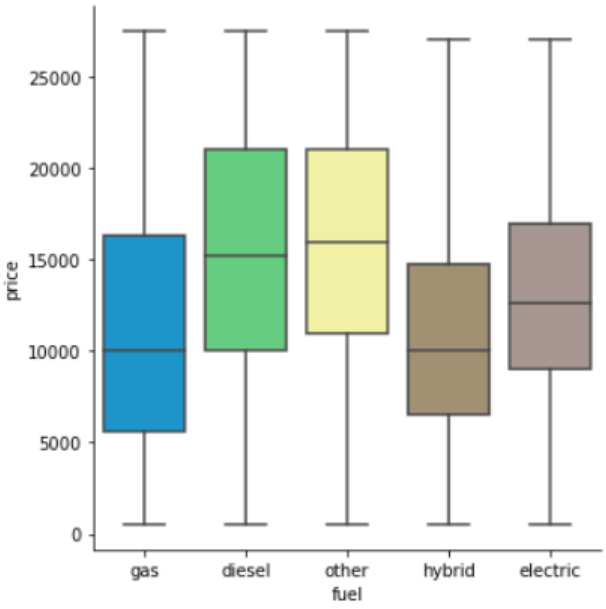


Figure 11. Fuel vs price

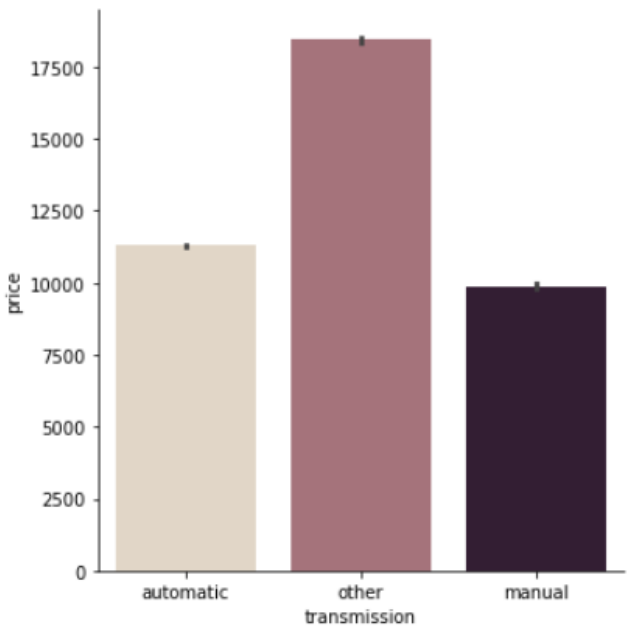


Figure 12. Transmission vs price

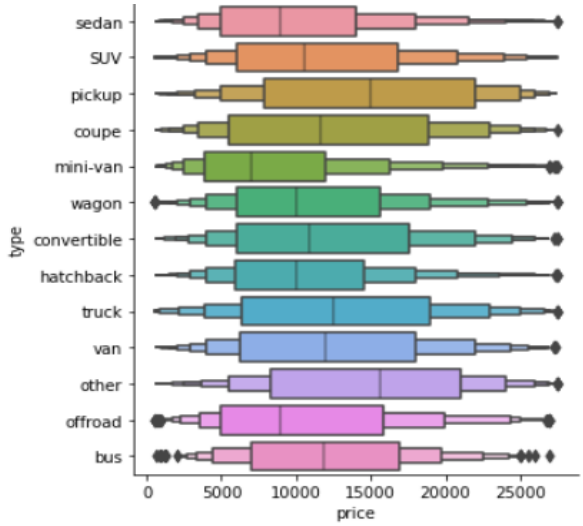


Figure 13. Vehicle type vs price

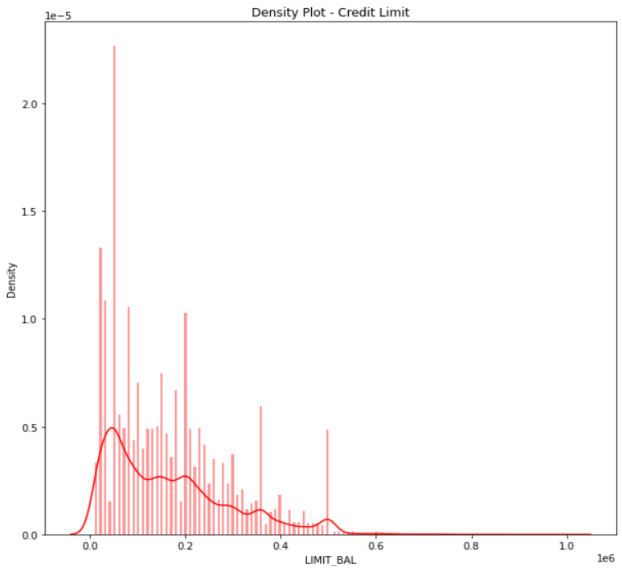


Figure 20. Density plot of Credit Limit

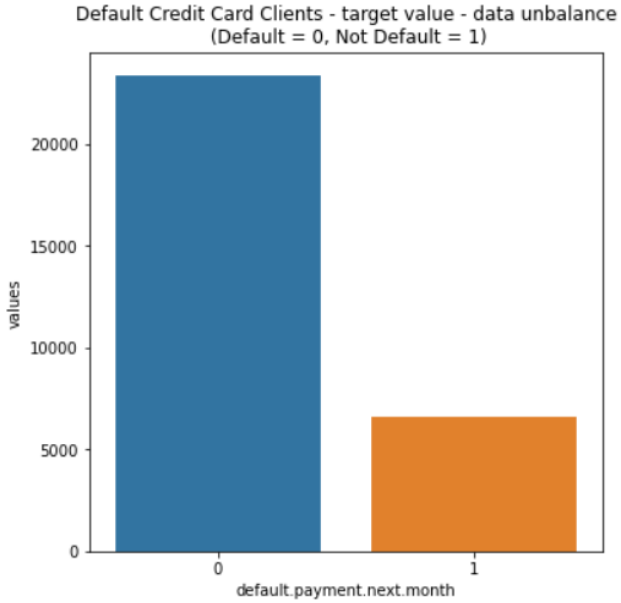


Figure 21. Target variable class distribution

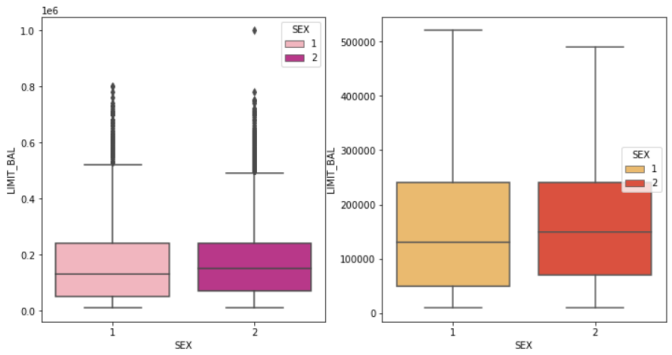


Figure 22. SEX distribution with respect to target variable

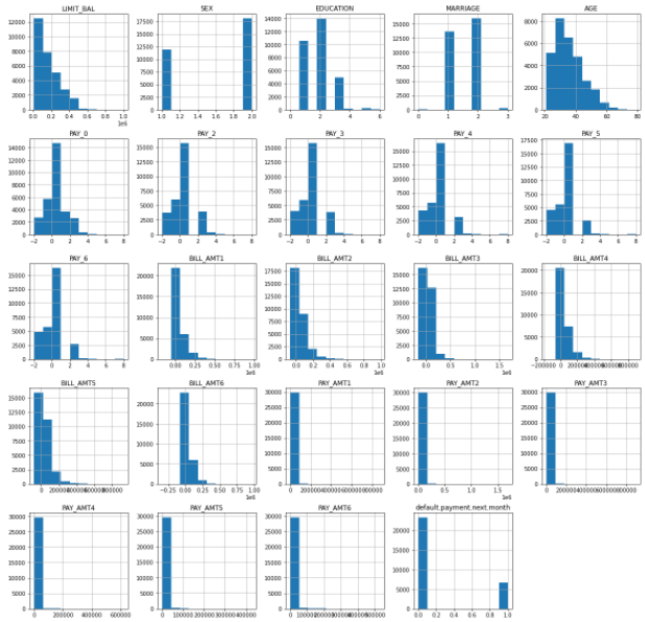


Figure 23. Distribution of variable

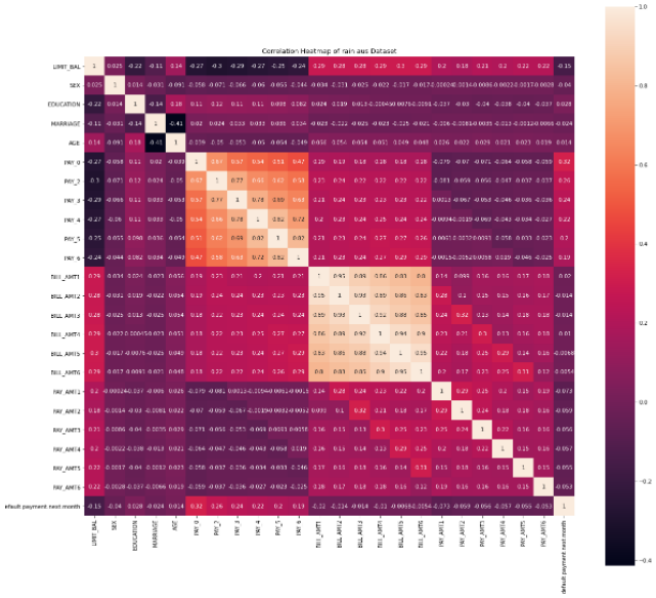


Figure 24. Correlation between variables

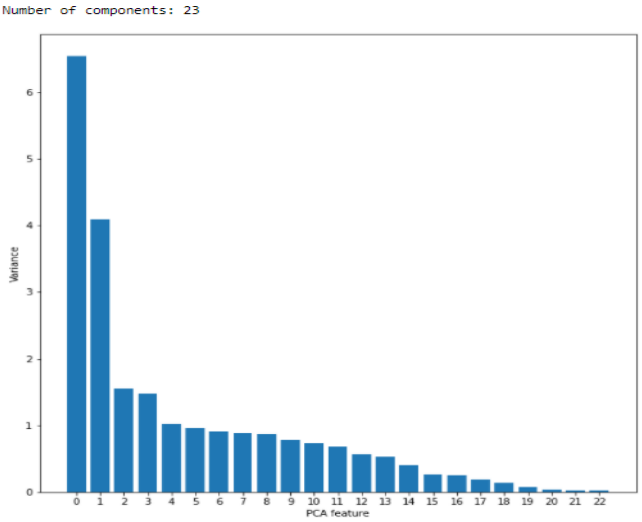


Figure 26. Factors vs variance