LOAN APPLICATION STATUS

Shobha



**Problem Statement:**

The Loan application status prediction data set that we are analysing is a data set which includes details of applicant who have applied for loan.

The dataset includes details like **Loan ID, Gender, Married, Dependents, Education, Self Employed, Applicant income, Co-applicant income, loan amount, loan amount term, and credit history.** We have data of some predicted loans from history**.**I have explored dataset and found a lot interesting facts about loan prediction.

The first part is going to focus on **Data analysis** and **Data visualization**. The second one we are going to see the about **Algorithm** used to tackle our problem.

The purpose of this analysis is to predict the loan eligibility process.

**Independent Variables:**

* Loan ID
* Gender
* Married
* Dependents
* Education
* Self Employed
* Applicant Income
* Co-Applicant Income
* Loan Amount
* Loan Amount Term
* Credit History
* Property Area

**Dependent Variable (Target Variable):**

* Loan Status

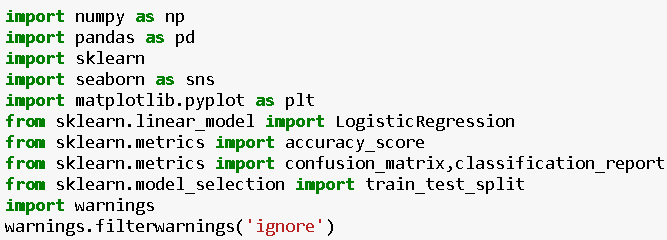
**Objective:**

The objective of the problem is to pick out which customer will be able to pay the loan and which customer is likely will not be able to pay the loan. We will create a classification model here and use algorithms like logistic regression, decision tree and random forest.

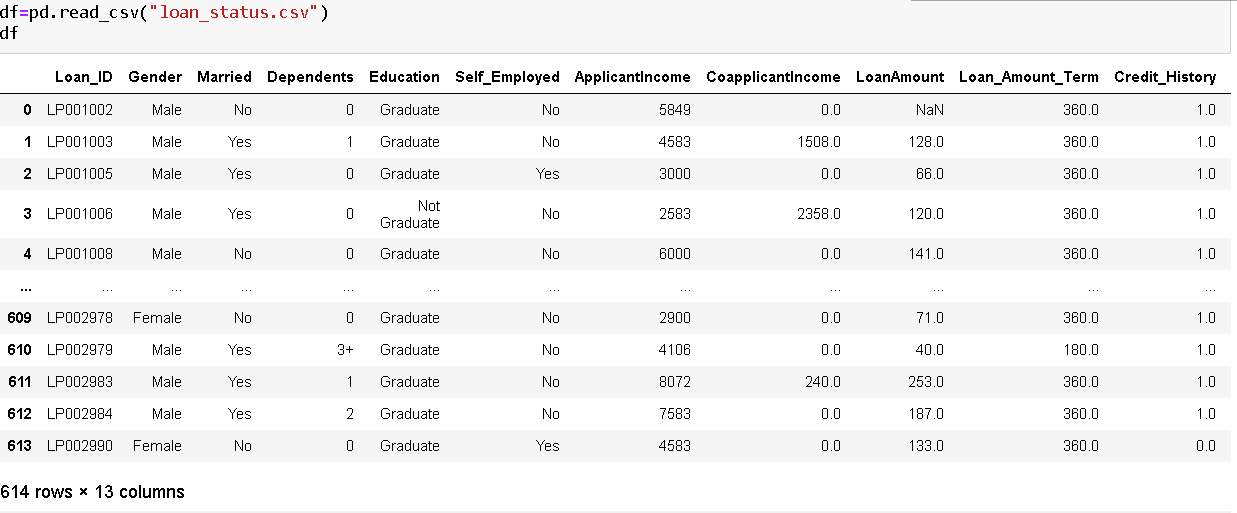
**Data pre-processing**

**Getting the system ready and loading the data**

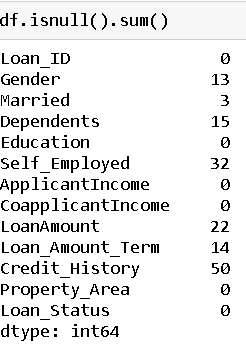
**We will be using Python for this course along with the below-listed libraries.**



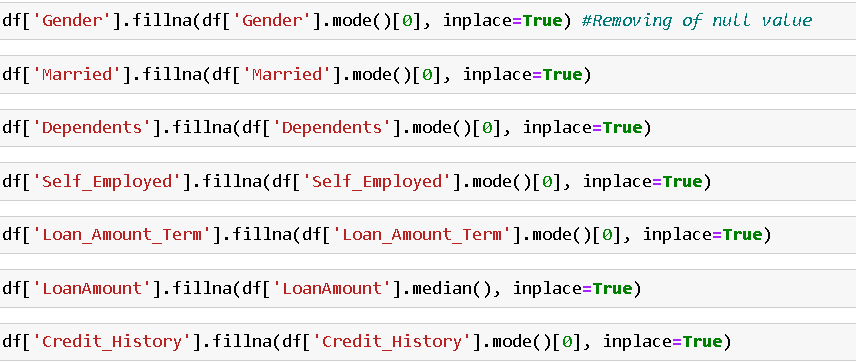
**Reading Data**



The dataset used in this project loan\_status.csv that contains 614 instances with 13 features.

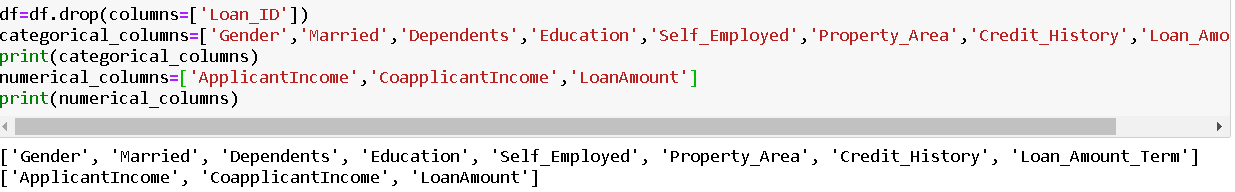


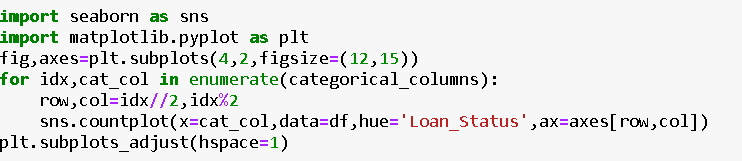
The data set has null value in Gender, Married, Dependents, Self-employed, Loan amount, Loan amount term and credit history. Let’s remove it according to columns type.

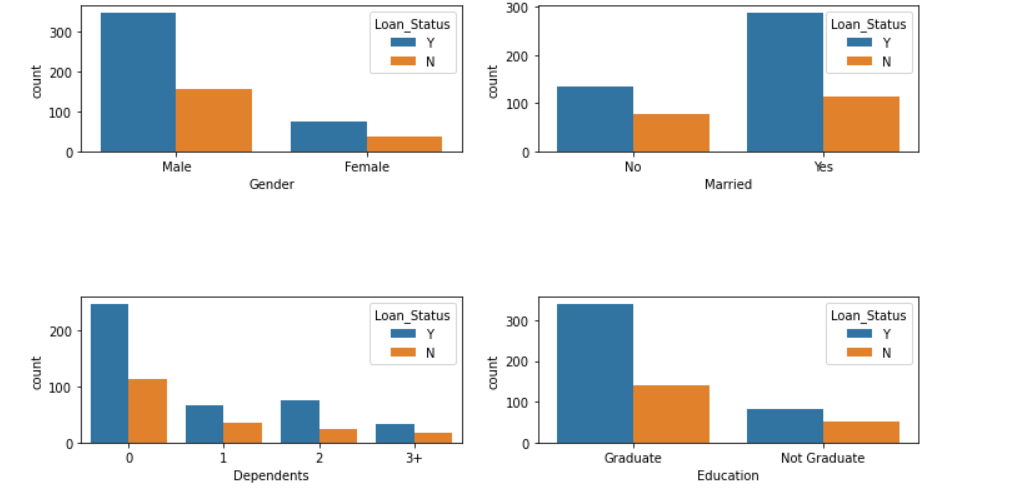


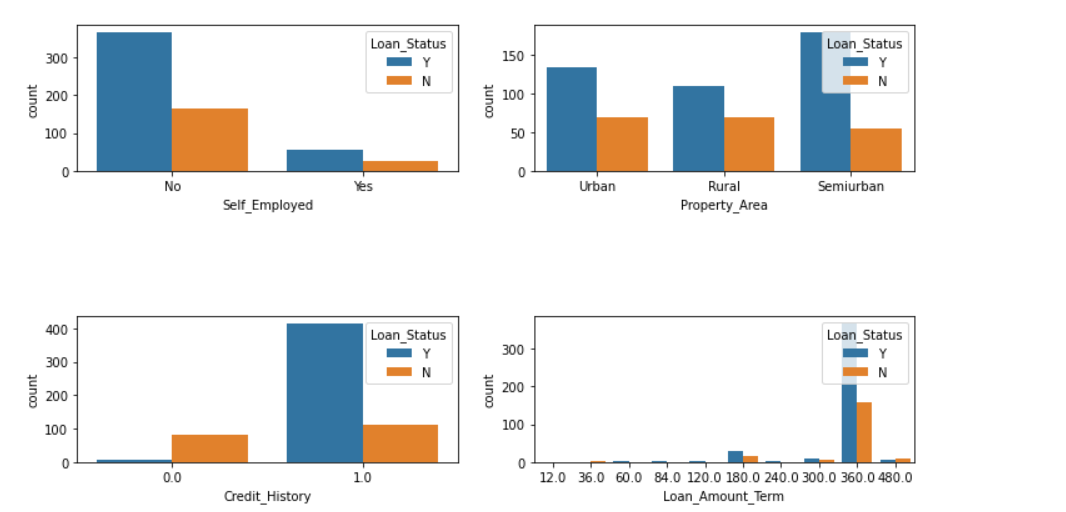
Dropping “loan Id” as it will not contibute any value while predicting the loan status.

**Visualizaing categorical and numerical columns**.

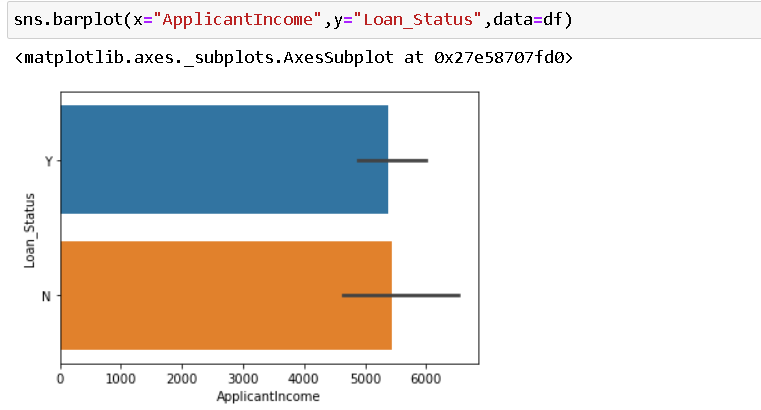


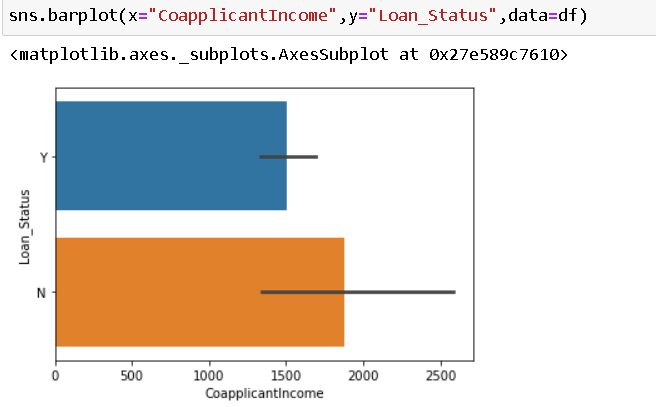


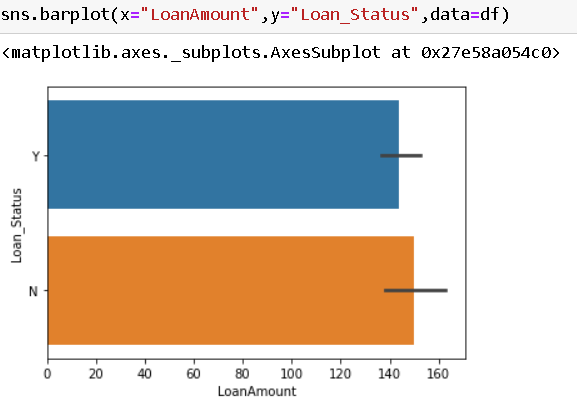




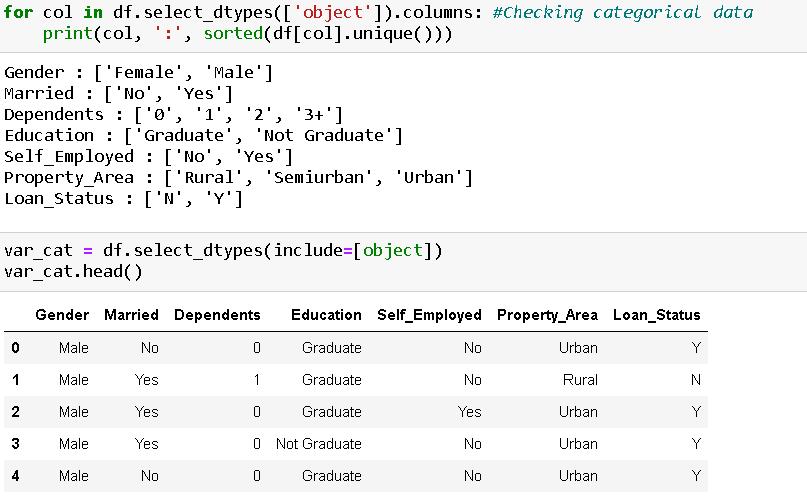
With the help of above graphics, we came to know that, Male applicant’s approval status is higher than women. Married couple got more approval of loan than unmarried. Less dependency lead to more chances of loan approval. Graduate applicants have higher proportion of loan approval. Self-employed applicant has lessor proportion of loan approval than other. Loan approval rate is higher for semiurban properties than urban and rural. Good credit history application has more chances to get the loan approval. Majority of approval got loan for 360 days.



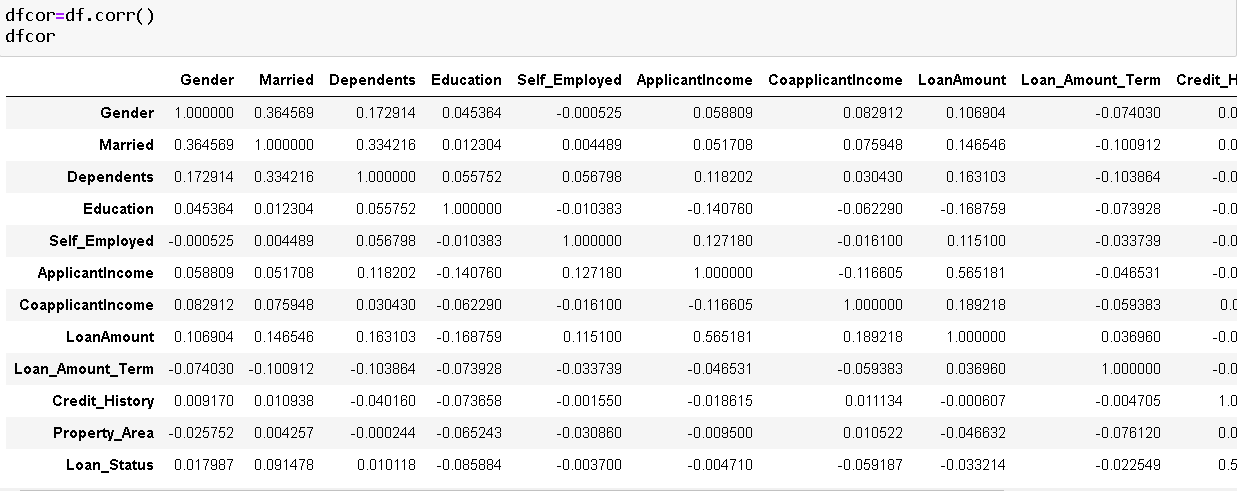


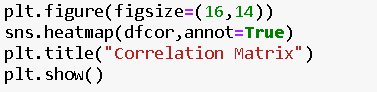


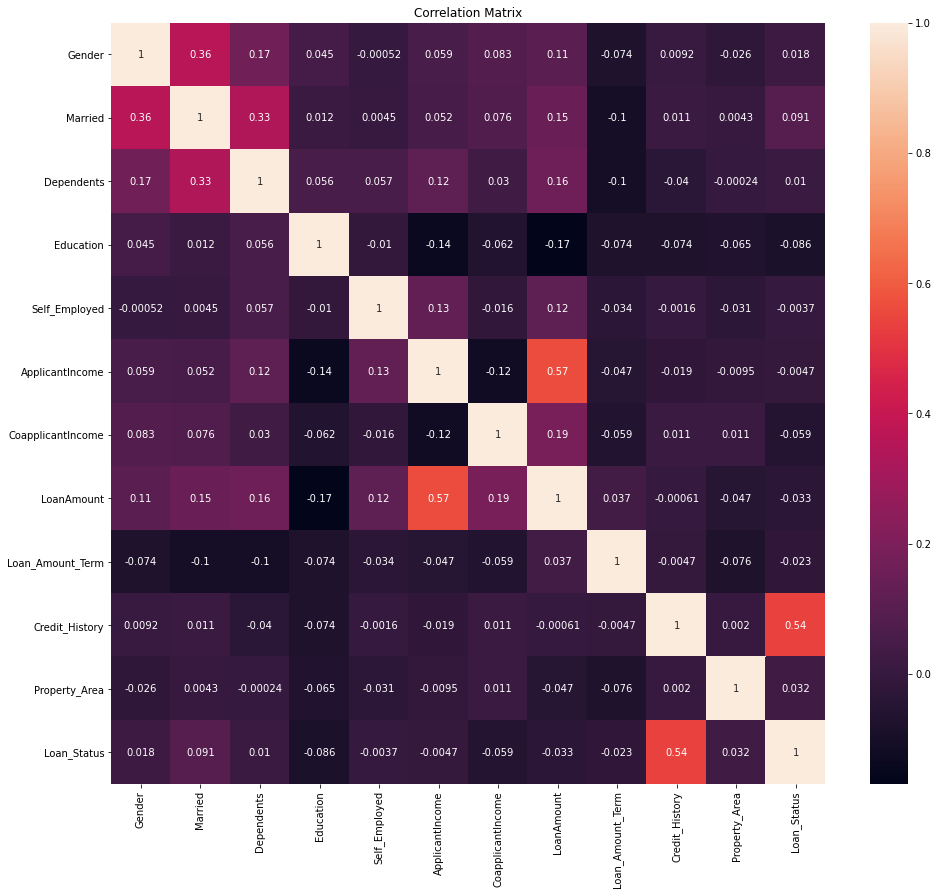
The above numeric graphics shows that there is not much correlation between loan status and applicant income, co-applicant income and loan amount





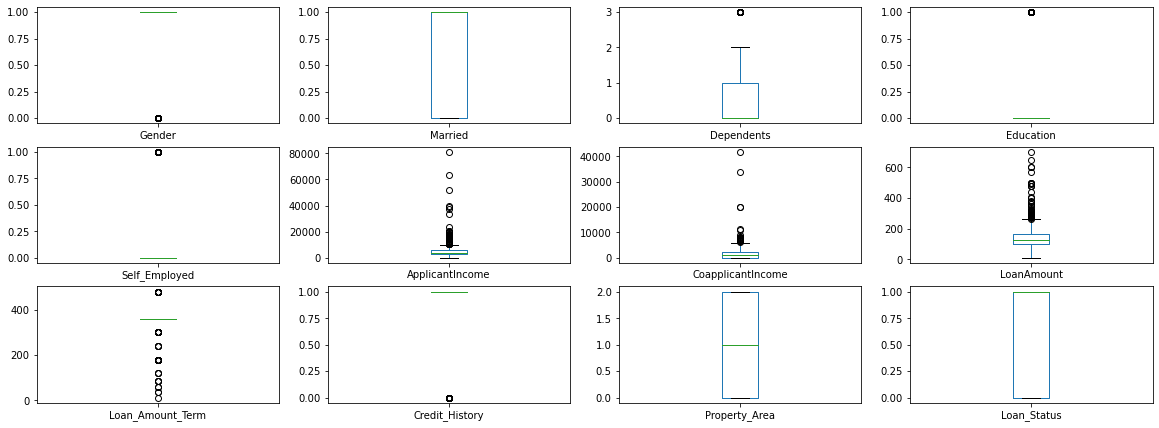




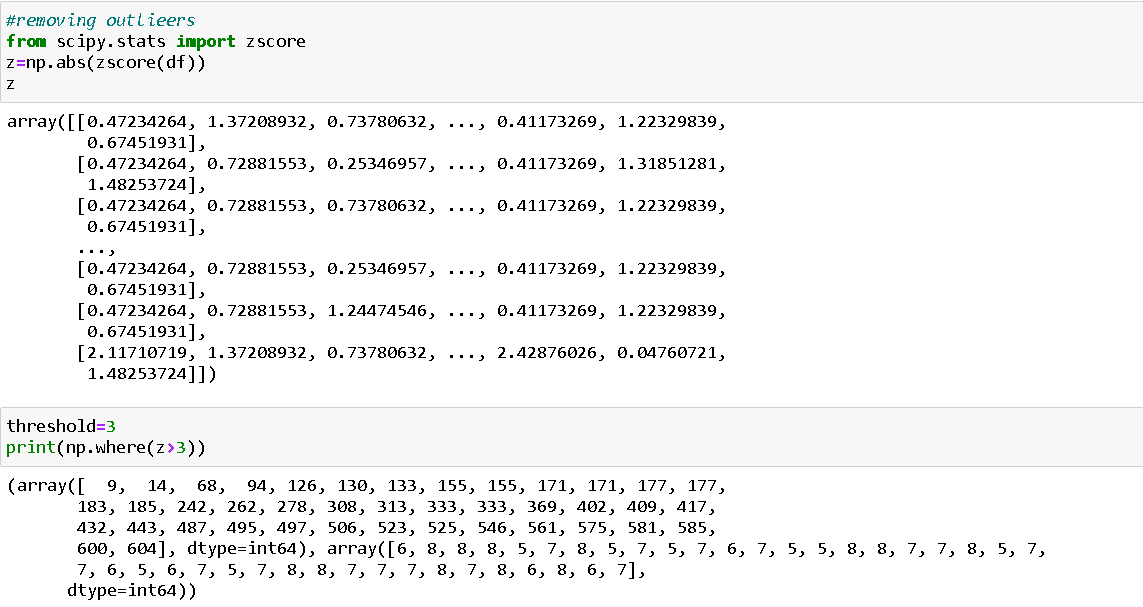


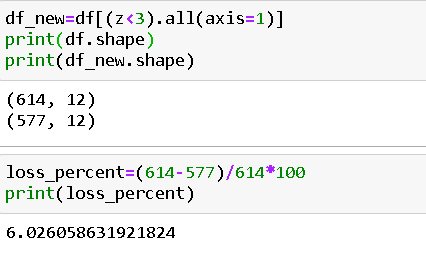
some of the variable like Credit History, Married, Property Area, Gender, Dependents have moderate correlation with target variable and some of the variable like Self-Employed, Applicant Income, Loan Amount Term, Loan Amount, Co-Applicant Income and Education have negative correlation with target variable



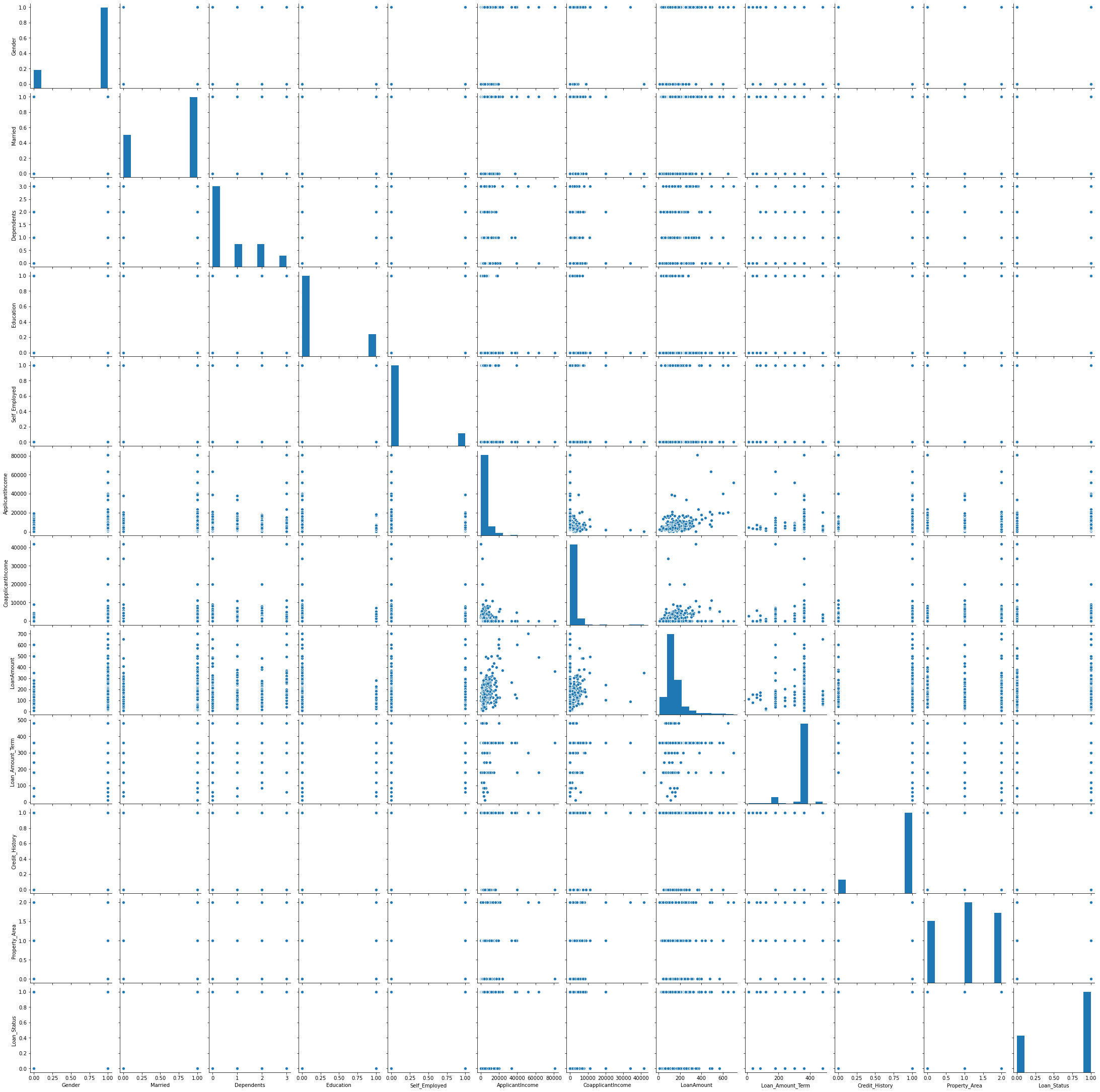


Applicant Income, Co-Applicant Income, Loan amount term columns have outliers. We will remove it with the help of Zscore

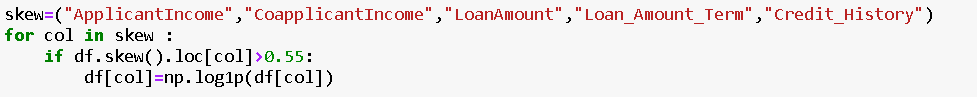




Checking and Removing of skewness.



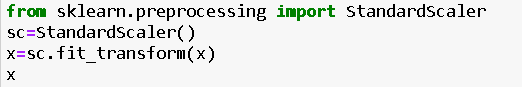
most of the columns has either right skewed or left skewed. Gender, married, loan amount term, credit history, property area and loan status has negatively skewed and dependents, education, self-employed applicant income Co-Applicant income and loan amount have positively skewed.



Separating Target from the feature for training

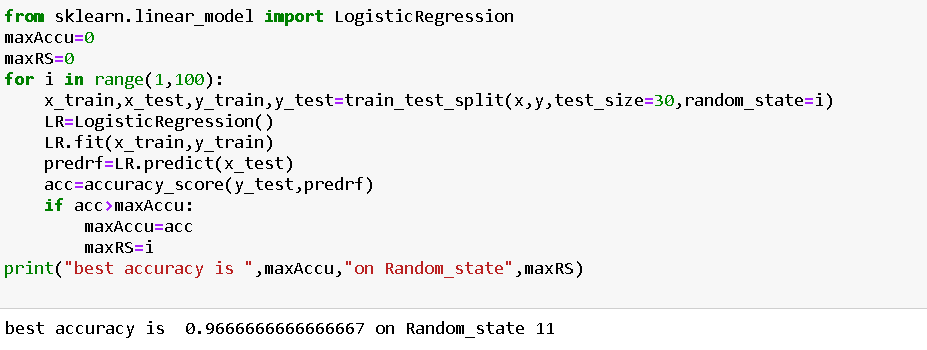


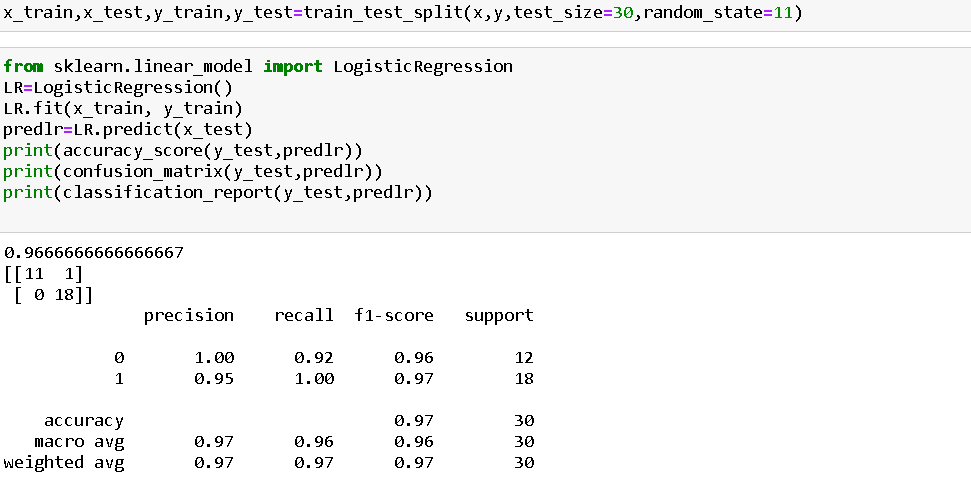
**Now We will work on Scalling, Training, Testing ,Validating and Hyper Tunning the Model**.

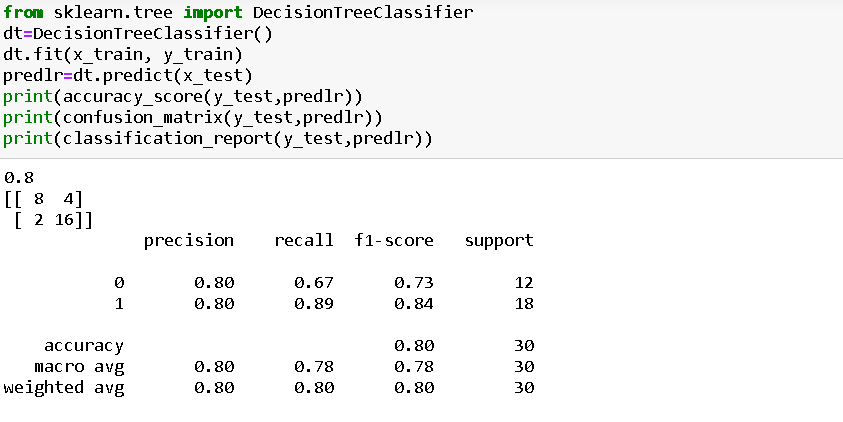


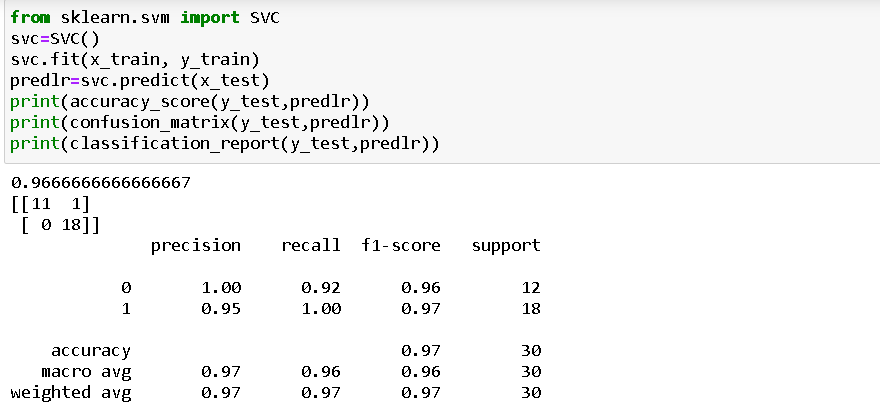
To begin, let’s split the dataset into training and test sets using 70/30 split; 70% of data will be used to train the model and the rest 30% to test the accuracy of the model. Then we can up sample the minority class, in this case the positive class.

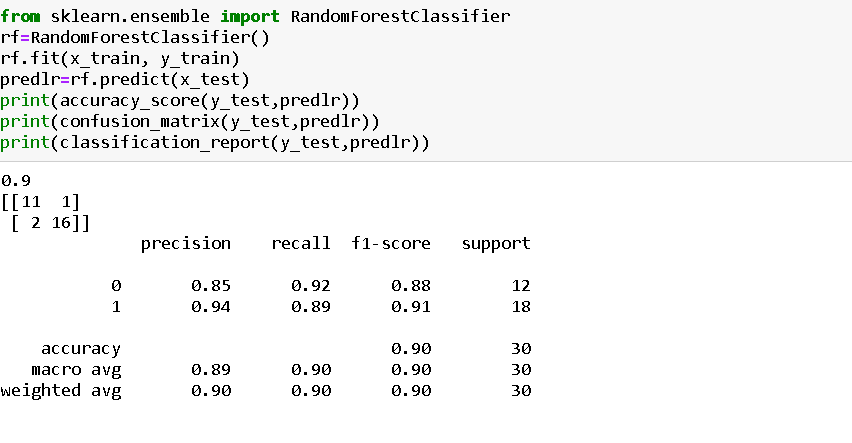
After partitioning and balancing, our data will make our model ready to be input of the machine learning models. We will train 5 different models: Logistic Regression, Random Forest, Decision Tree Classifier and SVC. In this step, we will start modifying model parameters, perform feature engineering and balancing data strategies to improve the performance of the models. Try with more trees in the Random Forest model, include new variables, penalize wrong predictions from the minority class until you beat the performance of our current best model.

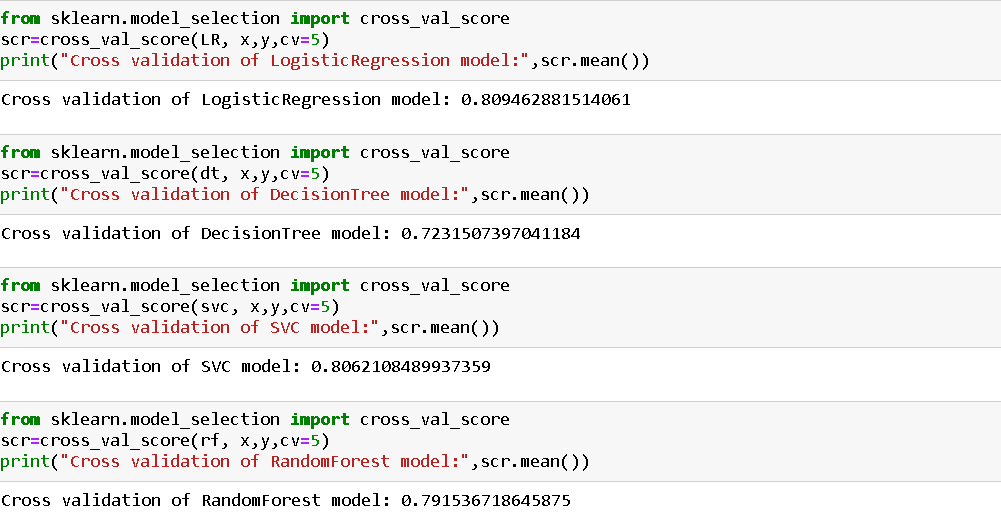


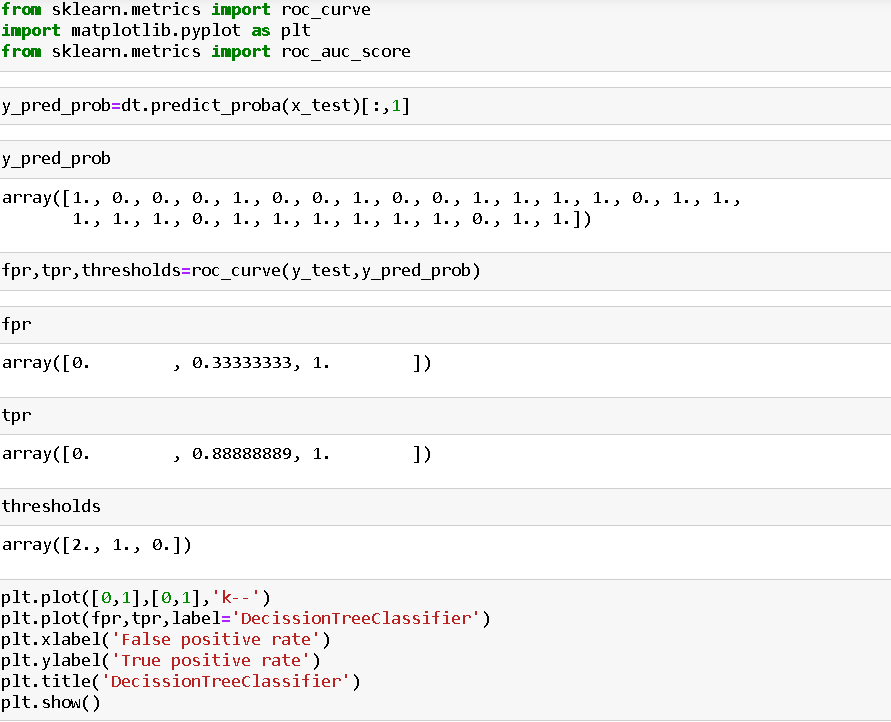


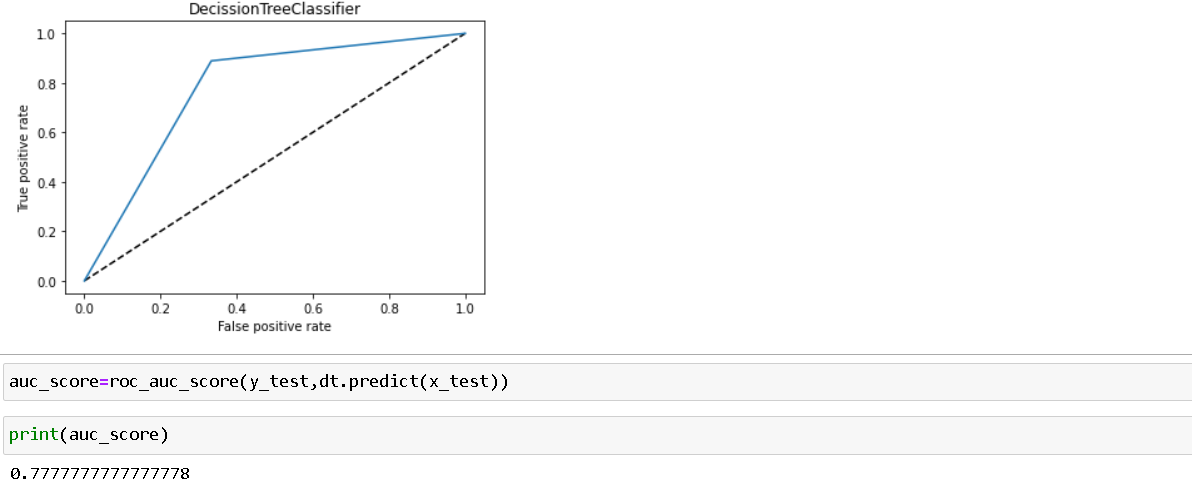




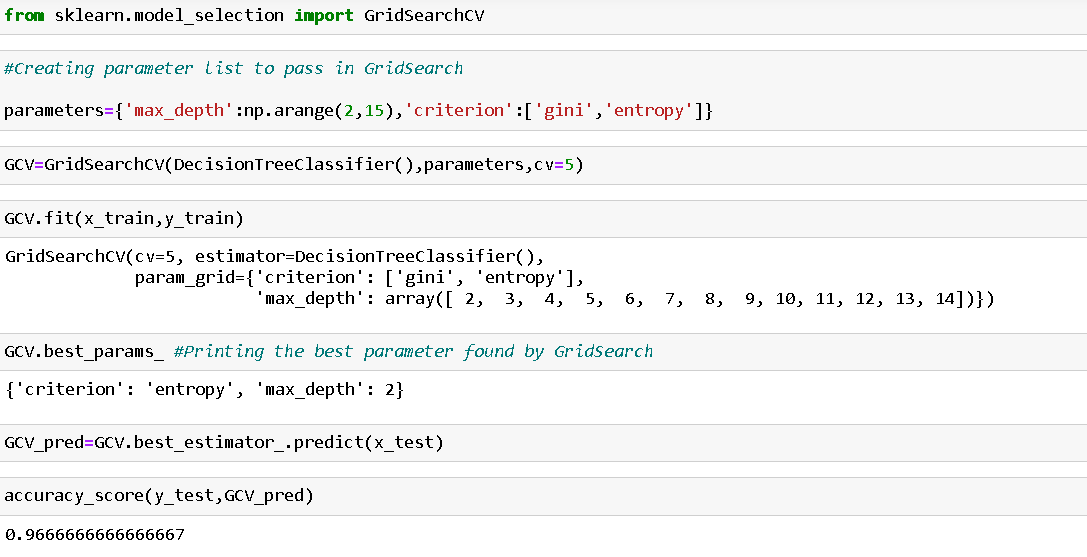




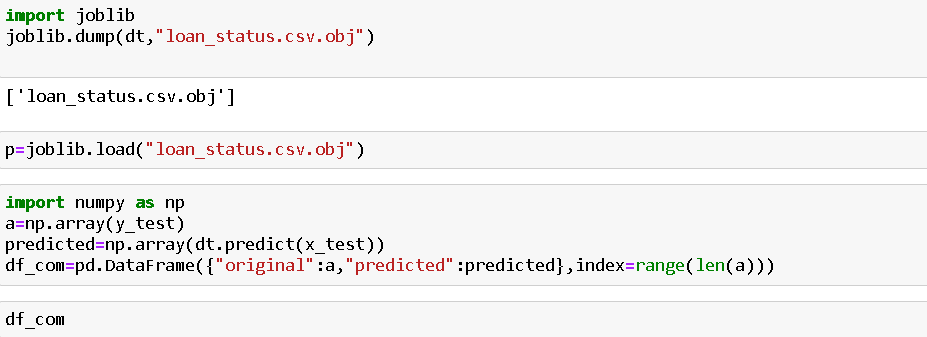


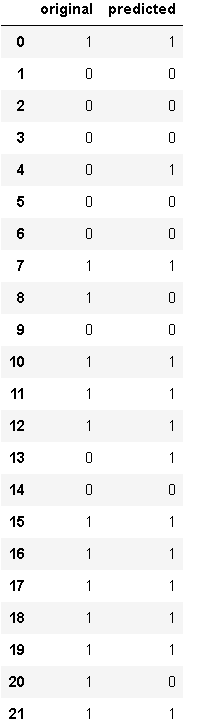


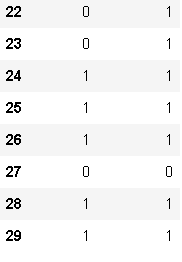
Finally, after testing our models with the test set, we concluded that best model was the DecisionTreeClassifier (dt). Now we will Hyper tune our model with the help of GridSearchCV to increase our model accuracy.



Saving Model







Conclusion

Nowadays, the loan business becomes more and popular, and many people apply for loans for various reasons. However, there are defaulter cases which results in huge financial loss. In this study, the dataset was cleaned first, and the exploratory data analysis and feature engineering were performed. The strategies to deal with both missing values and imbalanced data sets were covered. Then we propose four machine learning models to predict if the applicant could repay the loan, which are Random Forest, Logistic Regression, Support Vector Machine, and DecisionTreeClassifier. When tuning parameters, both Cross Validation and Grid Search Cross Validation methods are applied in different situations. It is noticed that the model was found which best fits the dataset with highest accuracy is the DecisionTreeClassifier model, and the model with highest AUC score is with DecisionTreeClassifier. As we expected, borrowers with higher annual income and higher credit scores are more likely to repay the loan fully; In addition, borrowers with lower interest rates and smaller instalments are more likely to pay the loan full.

Shobha

HR ANALYTICS - ATTRITION



**Abstract**

Employee attrition can become a serious issue because of the impacts on the organization’s competitive advantage. It can become costly for an organization. The cost of employee attrition would be the cost related to the human resources life cycle, lost knowledge, employee morale, and organizational culture. This study aimed to analyse employee attrition using logistic regression, Random Forest Classification, Decision Tree Classifier and SVC. The result obtained can be used by the management to understand what modifications they should perform to the workplace to get most of their workers to stay. We use python for data integration, exploratory data analysis, data preparation, logistic regression, Random Forest Classification, Decision Tree Classifier and SVC model evaluation, and visualization. The study has five steps: (1) data collection and business understanding, (2) data pre-processing, (3) exploratory data analysis, (4) model selection and training, and (5) test and evaluation of the model. The result of the study found eleven variables as key driving factors for employee attrition.

**Introduction:**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

Under HR analytics project we are analysing that what are the important factors which influenced attrition rate in an organization which results high cost to an organization. High attrition become a problem to any business. Our Study is based on 35 factors which can be a reason for high employee attrition.The input dataset is an Excel file with information about 1470 employees. For each employee, in addition to whether the employee left or not (attrition), there are features such as age, employee role, daily rate, job satisfaction, years at the company, years in current role, etc.

**The steps we will go through are:**

**Data Pre-processing**: Pre-Process the data to suit them with the analysis method.

The Pre-Processing may involve cleaning up the data, transforming the data, or creating new variables that may bring useful information for the analysis steps.

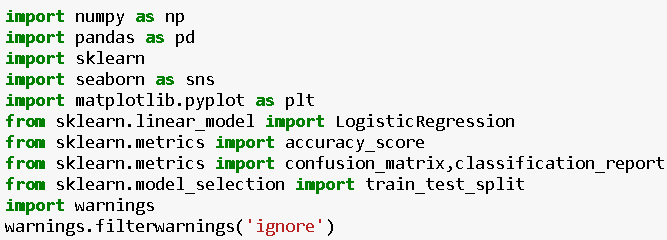
**Exploratory Data Analysis (EDA)**: This step creates textual and visual summaries of the dataset that highlight some characteristics of the data.

**Model Selection and Training, Test and Evaluate the Model**: Evaluate the performance of the proposed model

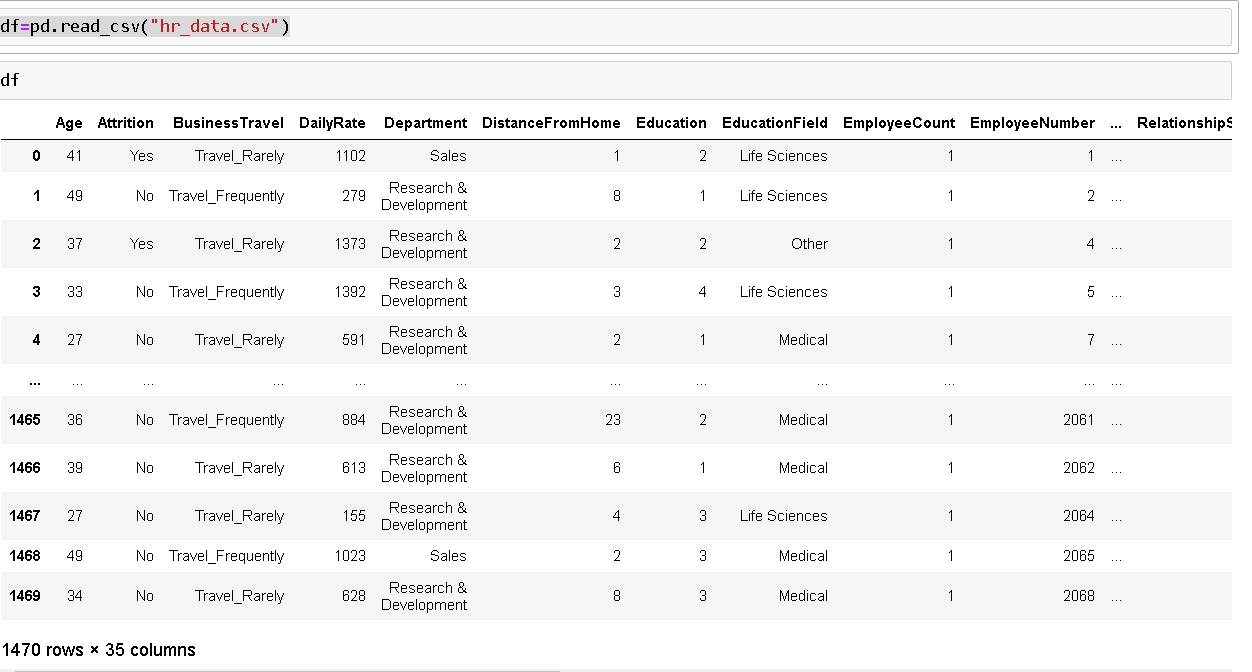
Data Pre-Processing

Getting the system ready and loading the data

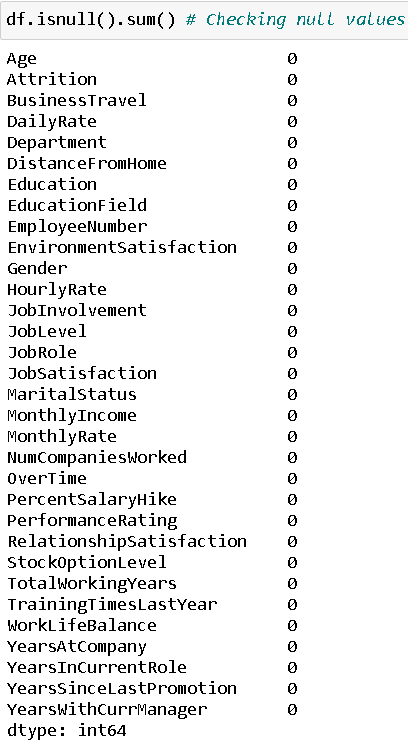
We will be using Python for this course along with the below-listed libraries.



**Reading Data**

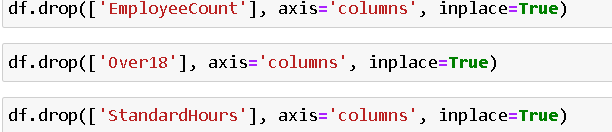


**Below are the 35 features of our dataset**: Age, Monthly income ,Attrition, Monthly rate ,Business travel, Number of previous employers, Daily rate, Over 18, Department ,Overtime, Distance from home, Per cent salary hike, Education ,Performance rating ,Education field ,Relations, satisfaction ,Employee count ,Standard hours ,Employee number ,Stock option ,level Environment satisfaction ,Total working years ,Gender ,Training times last year, Hourly rate, Work-life balance, Job involvement, Years with company ,Job level ,Years in current role, Job role, Years since last promotion ,Job satisfaction, Years with current manager ,Marital status.

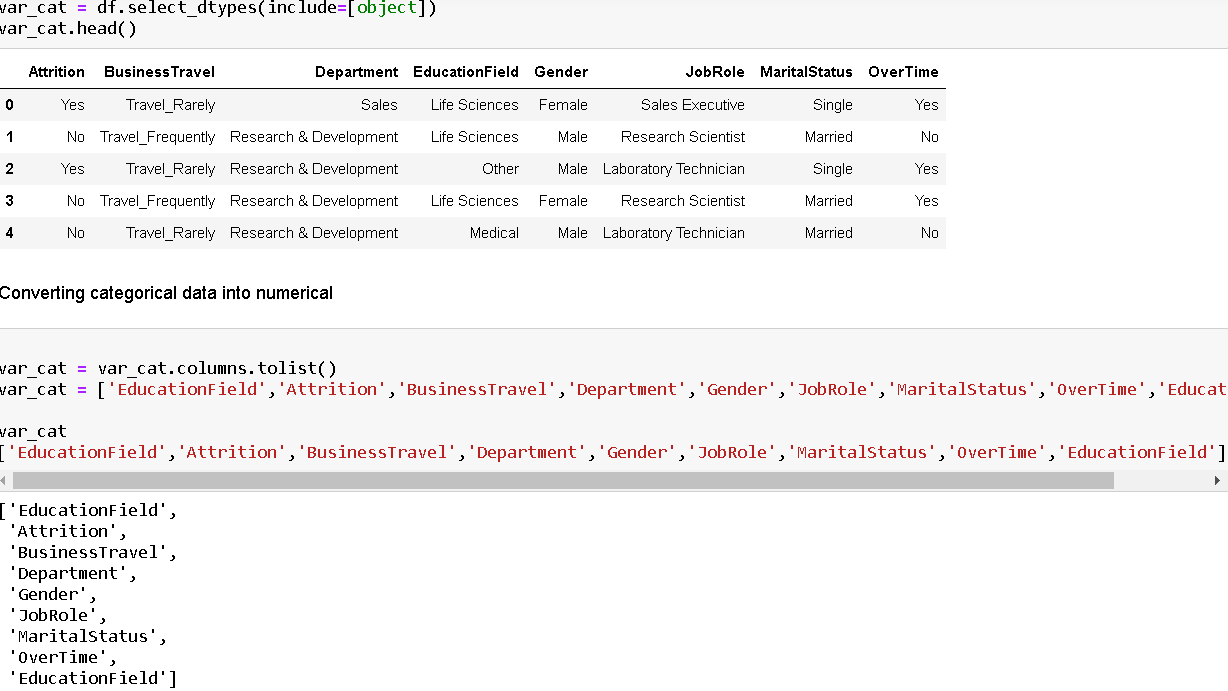


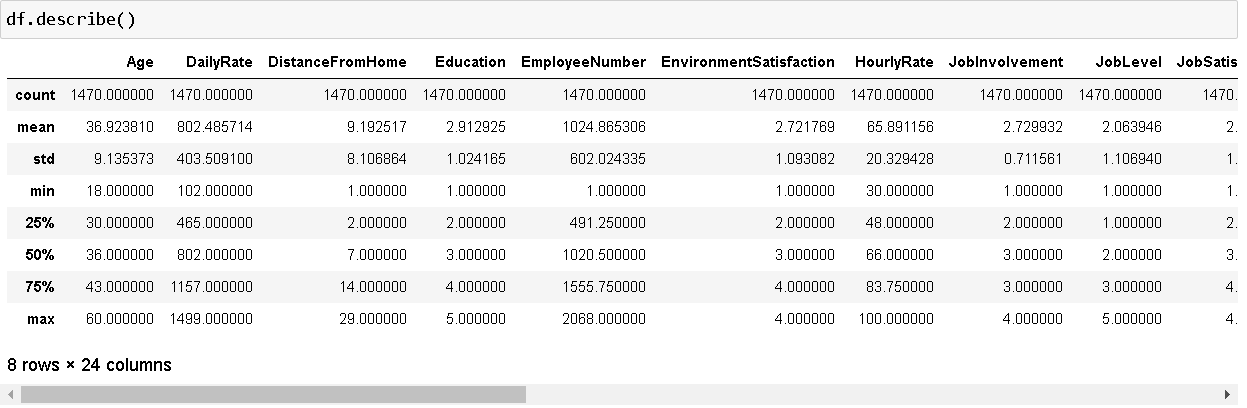
Good there is no null values in the given dataset.

We are removing Employee count, Over18 and Standard Hours as all of the columns have same values and hence not contributing much to predict in model accuracy.

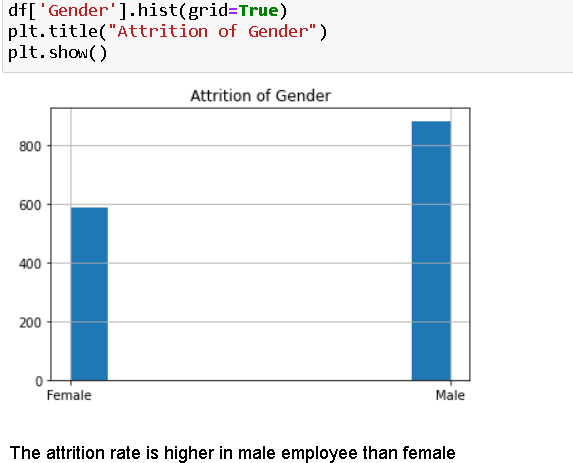


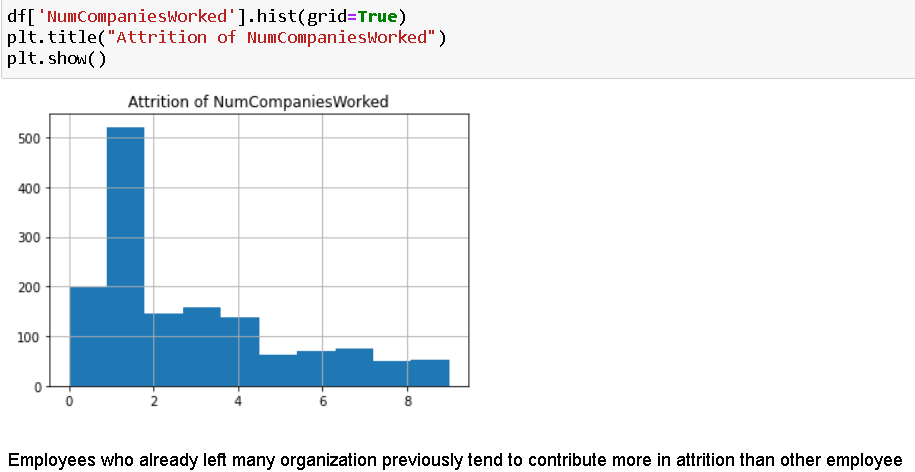
**Converting string values**

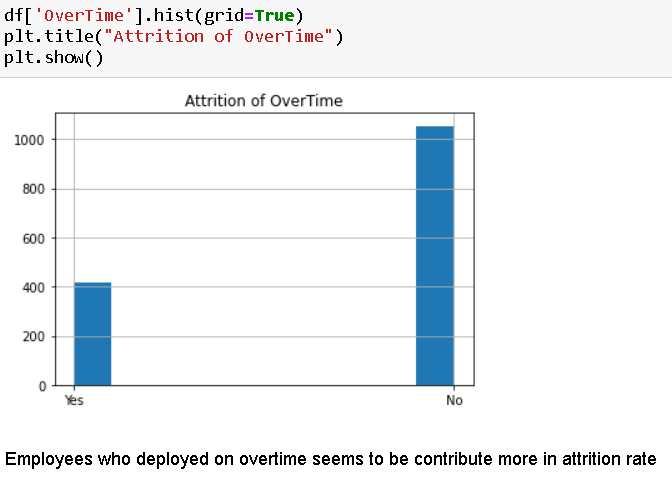


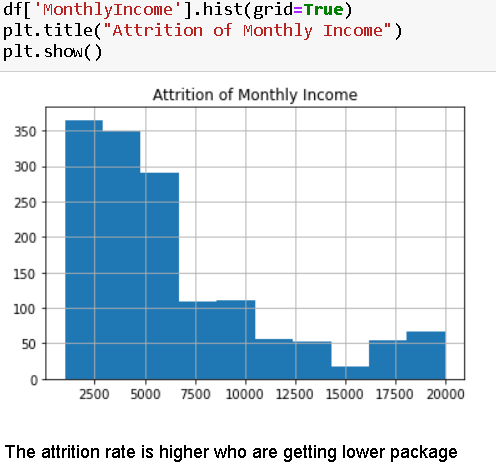
As per above statistics description there is no null value in given dataset. Most of the outliers falls in Total working year, training times last year, Years at company, years in current role, years since last promotion, years with current manager and education field. Standard deviation is on higher side in some of the columns. In some of the columns there is difference in mean and 50% and mean and standard deviation so the data is again skewed in columns.

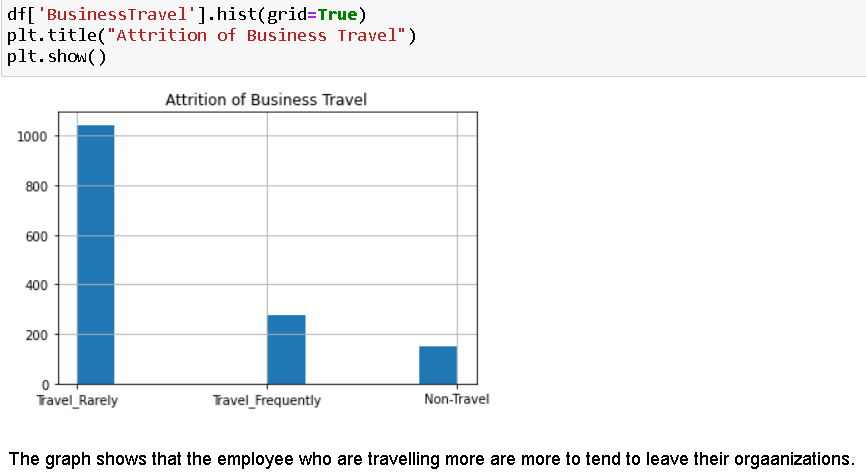
Data Visualization:

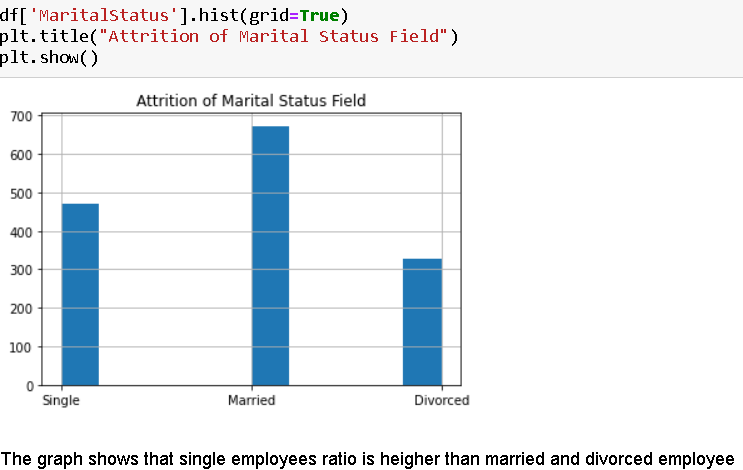


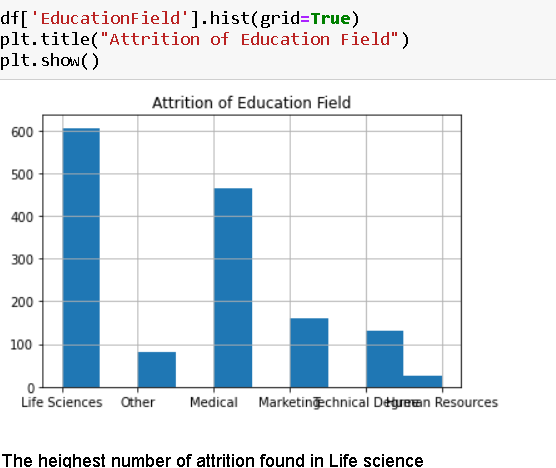


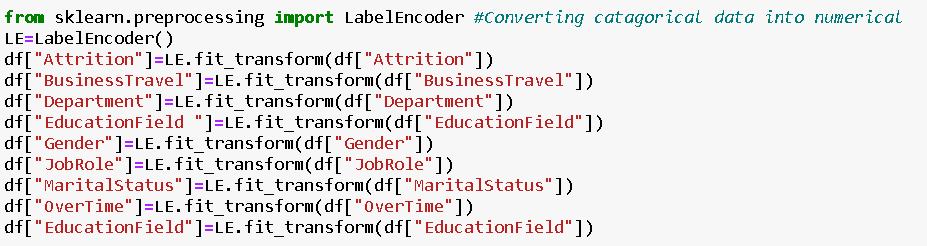


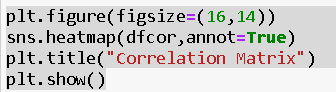


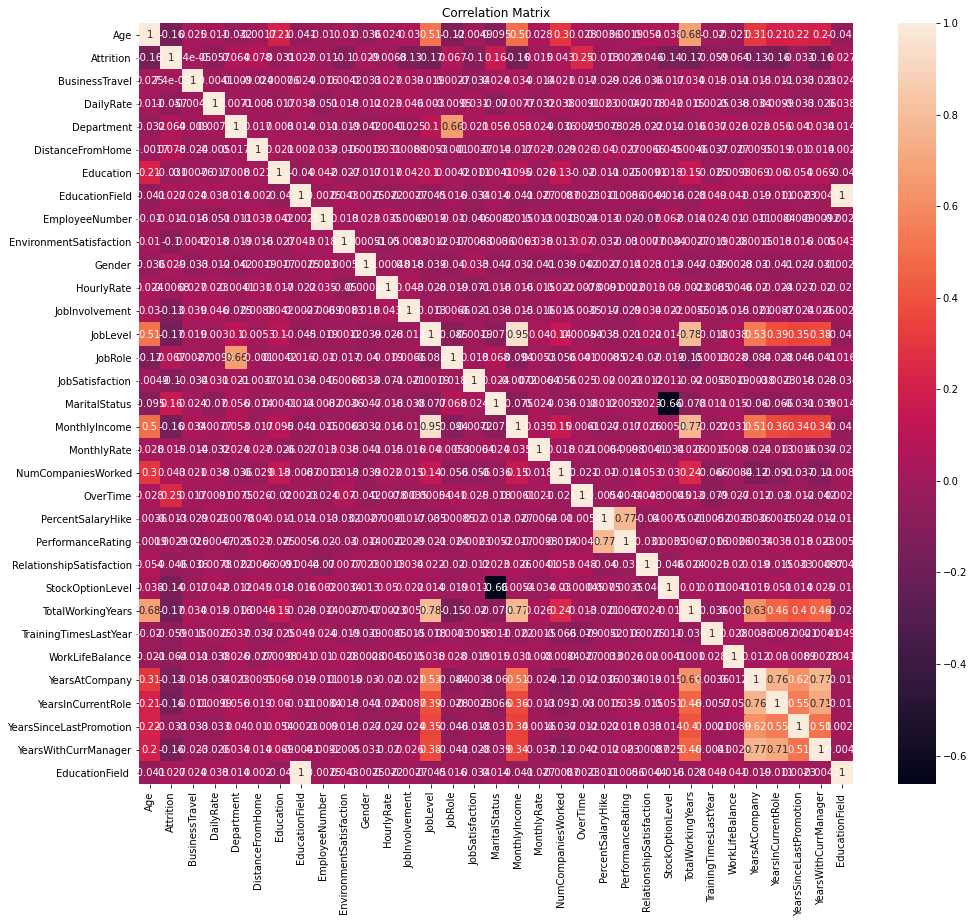






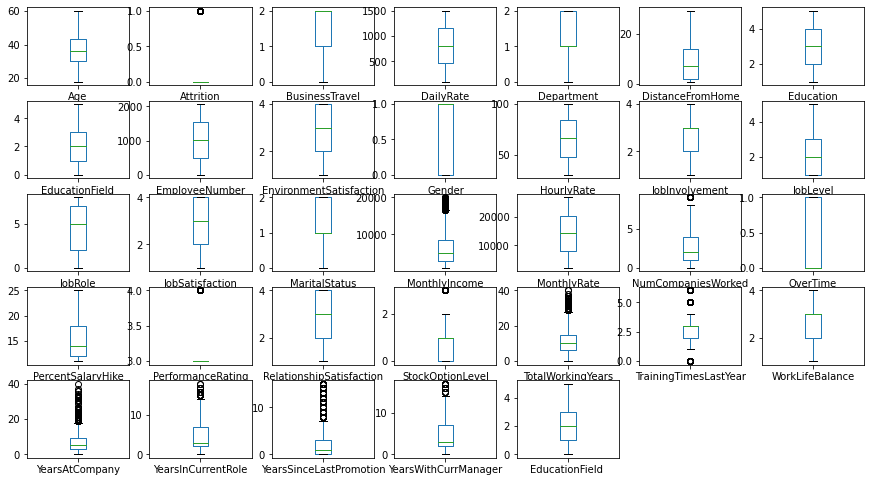






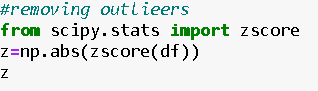
As per above graphics some of the variables are moderately correlated with Attrition and some of the variables are negatively correlated with attrition.

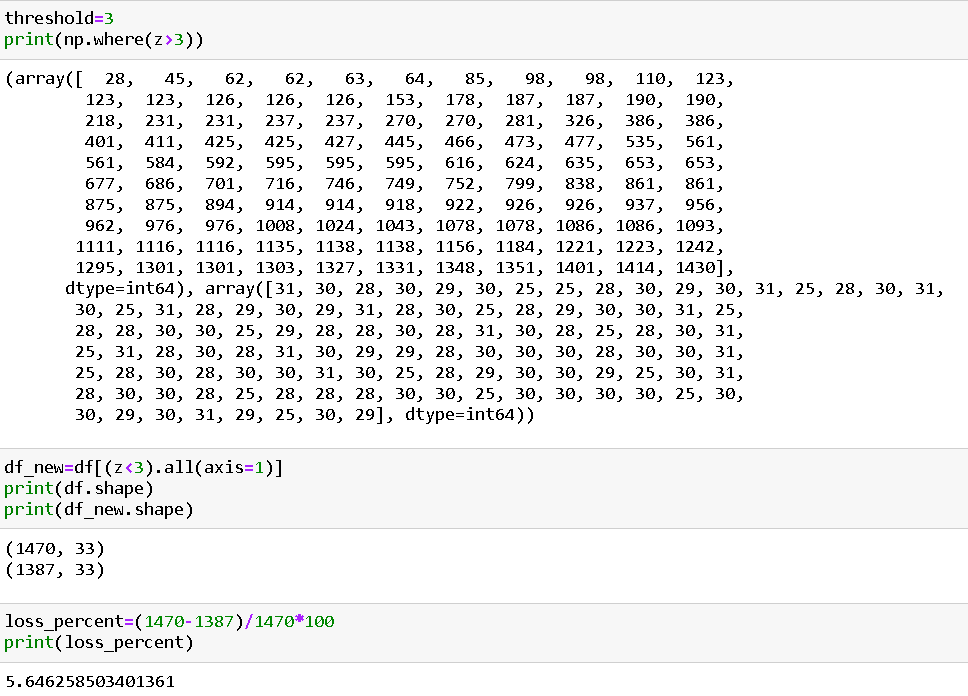
Checking outliers:



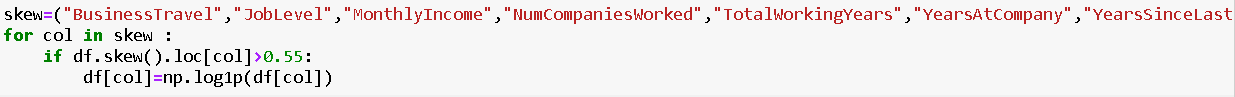
**The columns**

Monthly Income, Num companies worked, Stock option level, Total working year, Training times last year, Years at company, years in current Role, years since last promotion, years with current manager and education field has outliers. Let’s remove it to increase the model performance

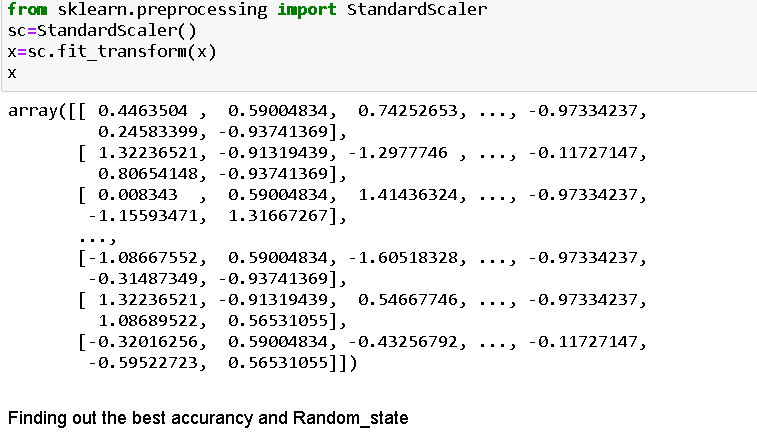




**Removing Skewness**



Now We will work on Scaling, training, testing, validating and hyper tunning the Model.

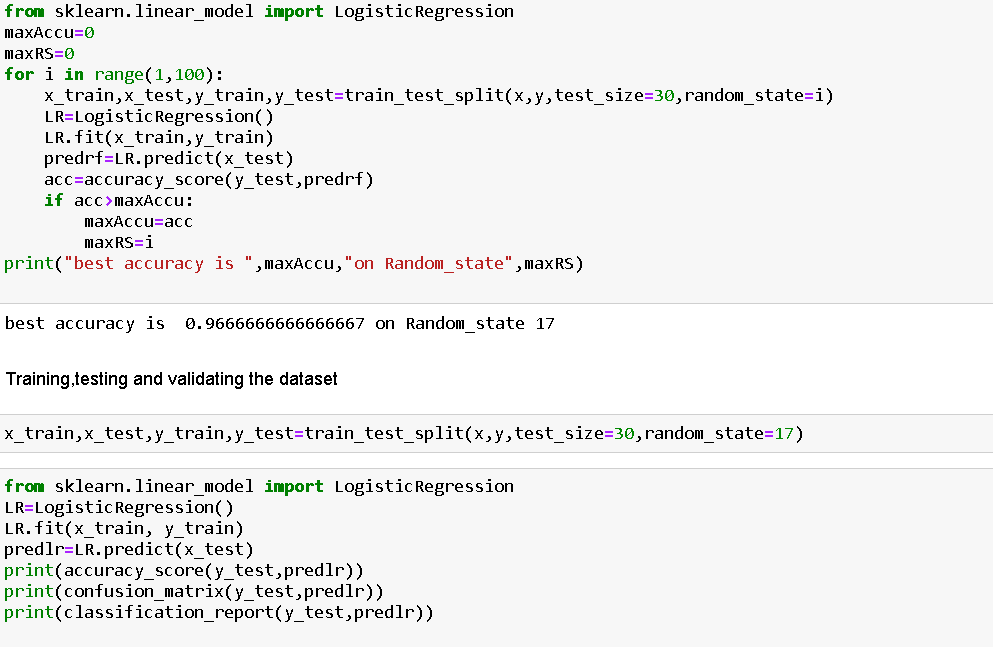


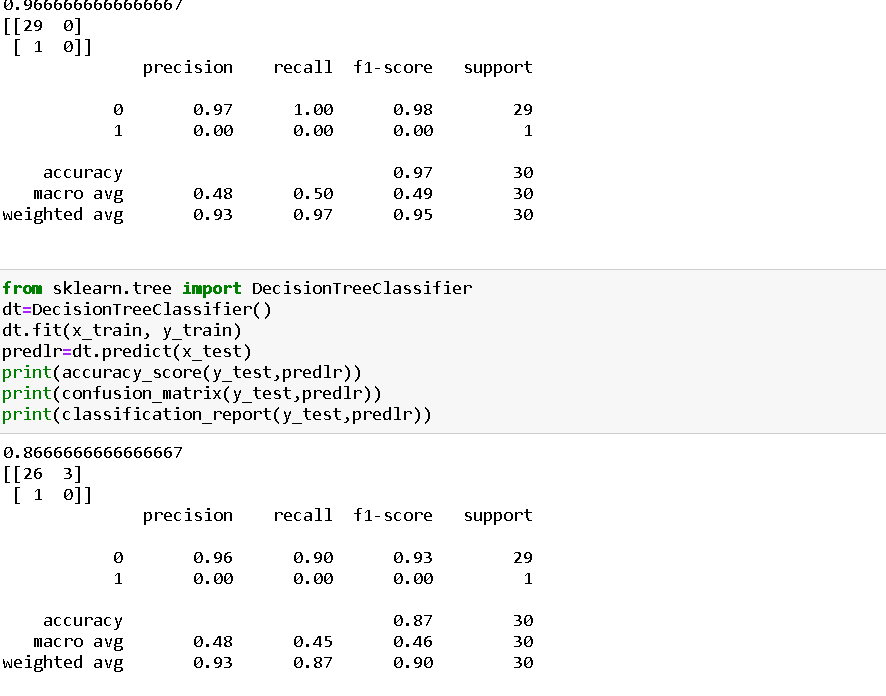
**Model Building**

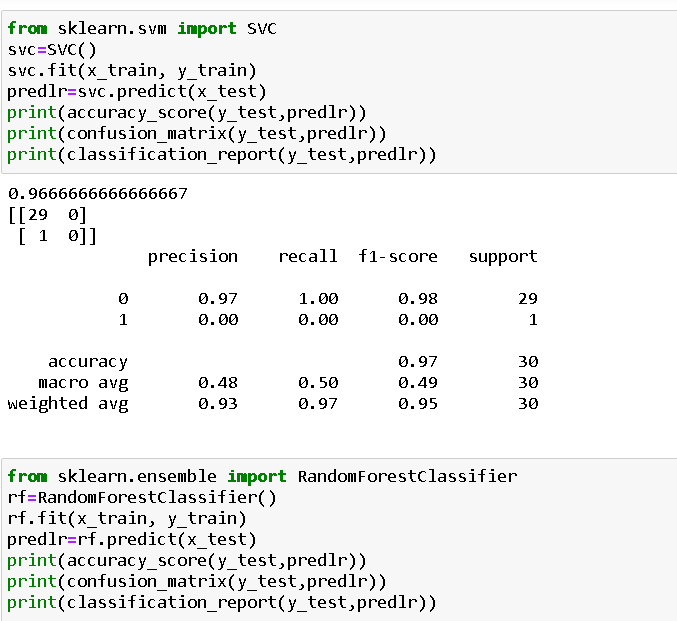
The modelling process consists in selecting models that are based on various machine learning techniques used in the experimentation. In this case various predictive models were used such as those based on decision tree, Random forest, logistic regression and SVM. The goal is to identify the best classifier for the analysed problem. Each classifier must therefore be trained on the featured set and the classifier with the best classification results is used for prediction. The classification algorithms taken into consideration are: • Logistic Regression classifier, • Decision tree classifier, • Random forest classifier, • Support Vector Machines (SVM) classification. After identifying the objectives and adequately preparing and analysing the dataset to be used, we proceeded with the design of the prediction model to identify employees that would potentially leave the company. In the construction phase of a model that implements a supervised learning algorithm, it was necessary to have a training-set available that consisted of instances of an already classified population (target), in order to train the model to classify new observations, which will constitute the test-set (in which the attribute representing the class was missing). Then, the model must be trained on a consistent number of observations in order to refine its prediction ability. The precision of the machine learning algorithms increases with the amount of data available during training.

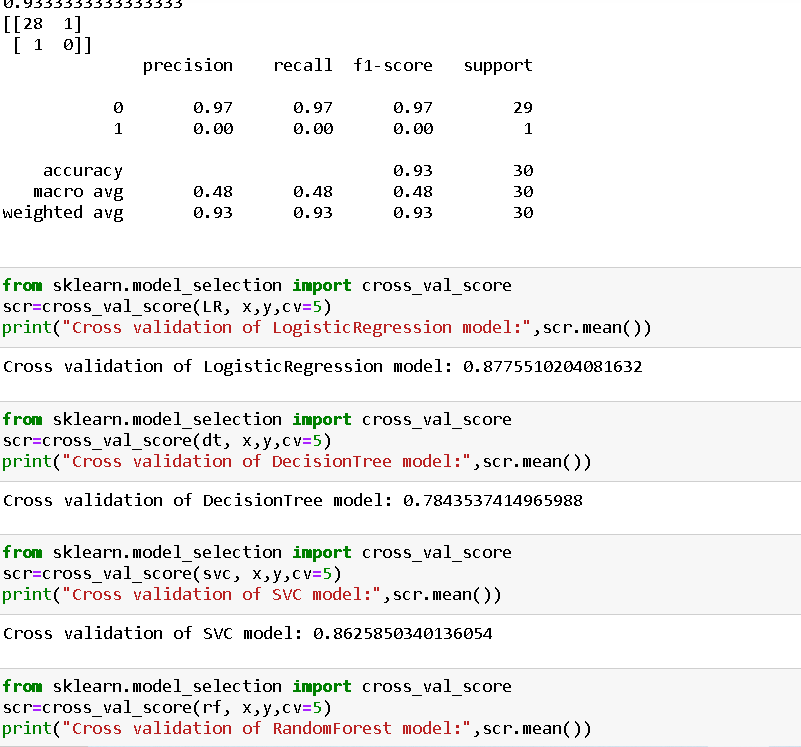
To begin, let’s split the dataset into training and test sets using 70/30 split; 70% of data will be used to train the model and the rest 30% to test the accuracy of the model. Then we can up sample the minority class, in this case the positive class.

In this step, we will start modifying model parameters, perform feature engineering and balancing data strategies to improve the performance of the models. Try with more trees in the Random Forest model, include new variables, penalize wrong predictions from the minority class until you beat the performance of our current best model.





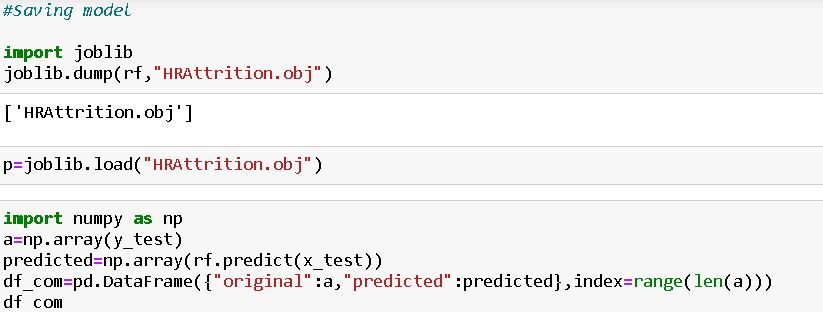


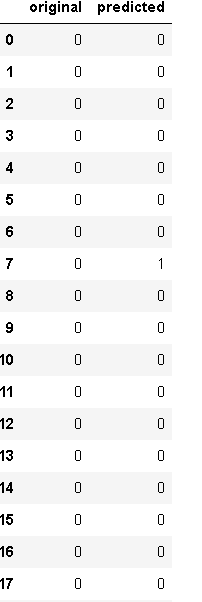


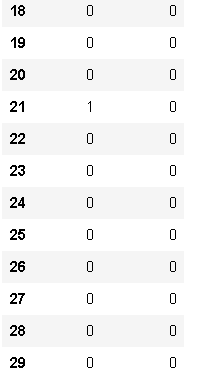
Model validation

Finally, after testing our models with the test set, we concluded that best model was the Random Forest (RF). Now we will Hyper tune our model with the help of GridSearchCV to increase our model accuracy.









**Wrapping up**

In this blog article we have detailed the various steps when implementing analytics use case in HR, employee attrition. We used the python to prepare the data, train different models, compare them and chose the best. With the model predictions, we created a model that would help any HR manager to retain the best talent by applying the correct strategies. This step-by-step blog article is just an example of what analytics can do for your business, and of how easy is to do it with the proper tool.