

PREPARED BY:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Pralad Kadel** Data Analyst Apprentice | |  | |  | | --- | | Chakupat-10, Lalitpur Nepal | | [coderush.com.np](http://coderush.com.np/) | |

Default Prediction Model

PROJECT REPORT

PREPARED FOR:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Shristi Shrestha** Associate Project Manager | |  | |  | | |  | | --- | | Chakupat-10, Lalitpur Nepal | | [coderush.com.np](http://coderush.com.np/) | |

## Introduction

To every bank, it is very important that the person taking loans from them pay back on time. Predicting the outcome of a loan is a recurrent, crucial and difficult issue in insurance and banking. Thus, knowing which clients are likely to default in advance can be very beneficial to them. So, this project is an attempt to create a machine learning model that can predict beforehand if there’s any chance of account default of a customer.

The objective of our project is to predict whether a loan will default or not based on borrowers’ income, loan amount, provided personal information and other factors.

The model is intended to be used as a reference tool for the client and his/her financial institution to help make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized.

## Key Aspects of the Assessment

## Data:

The data set that is being used for this project is a credit risk data set from the site Kaggle, with 12 features and around 32k records. Data includes loan details, e.g., amount, purpose, status, interest rate and loan grade and Customer details, e.g., age, income, employment, loan percent income, default history etc. Table below represents the data dictionary of final attributes that will be utilized in our model.

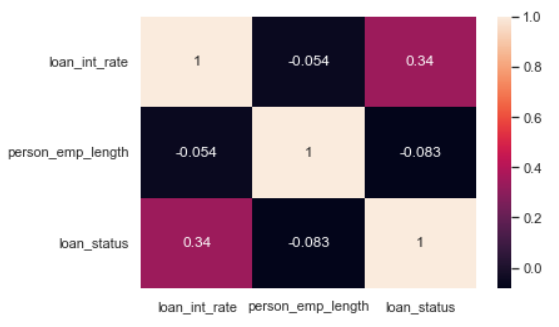
|  |  |
| --- | --- |
| Feature Name | Description |
| person\_age | Age of the customer |
| person\_income | Annual income |
| person\_home\_ownership | Home ownership status |
| person\_emp\_length | Employment length (in years) |
| loan\_intent | Intent of the loan |
| loan\_grade | Grade of the loan |
| loan\_amnt | Total loan amount |
| loan\_int\_rate | Interest rate |
| loan\_status | Status of the loan (default or not) |
| loan\_percent\_income | Percent income |
| cb\_person\_default\_on\_file | Historical default |
| cb\_person\_cred\_hist\_length | Credit history length |

Dataset link: <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

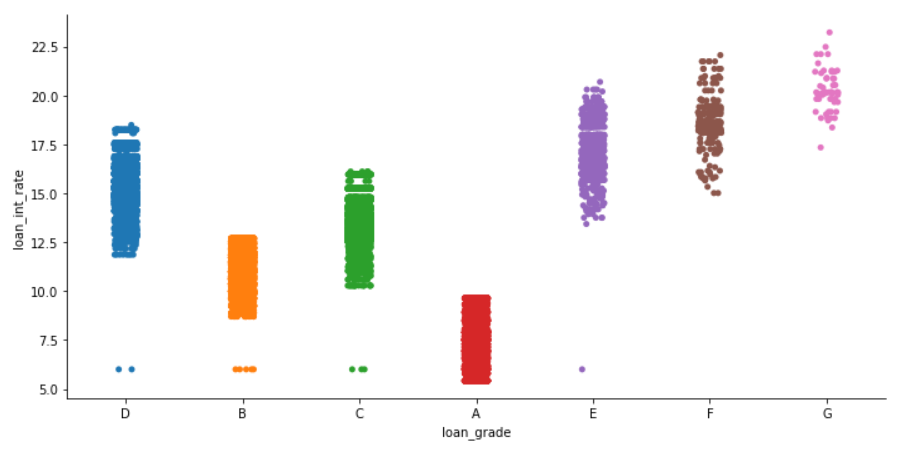
## Methods:

For the analysis of data, various types of descriptive statistics like measure of frequency, correlation, dispersion etc. are used so that we can present the data in a meaningful and understandable way allowing us a simplified interpretation and visualization of the data. We have used various univariate plots (like histogram, pie chart etc.) and multivariate plots (like scatter, categorical plots etc.) for the analysis of the data to understand their relationship with the loan default chances.

## Data cleaning and pre-processing:

Data cleaning includes handling missing values, removing outliers, removing irrelevant and duplicate data etc. Our dataset had two columns with missing values namely ‘person\_emp\_length’ and ‘loan\_int\_rate’.

*Fig: Illustration of correlation among missing values and loan status*

The person employment length column had only few missing values and since it didn’t have high correlation with the loan default status column, we dropped the missing values for this column. On the other hand, loan interest rate affected the default chance directly.

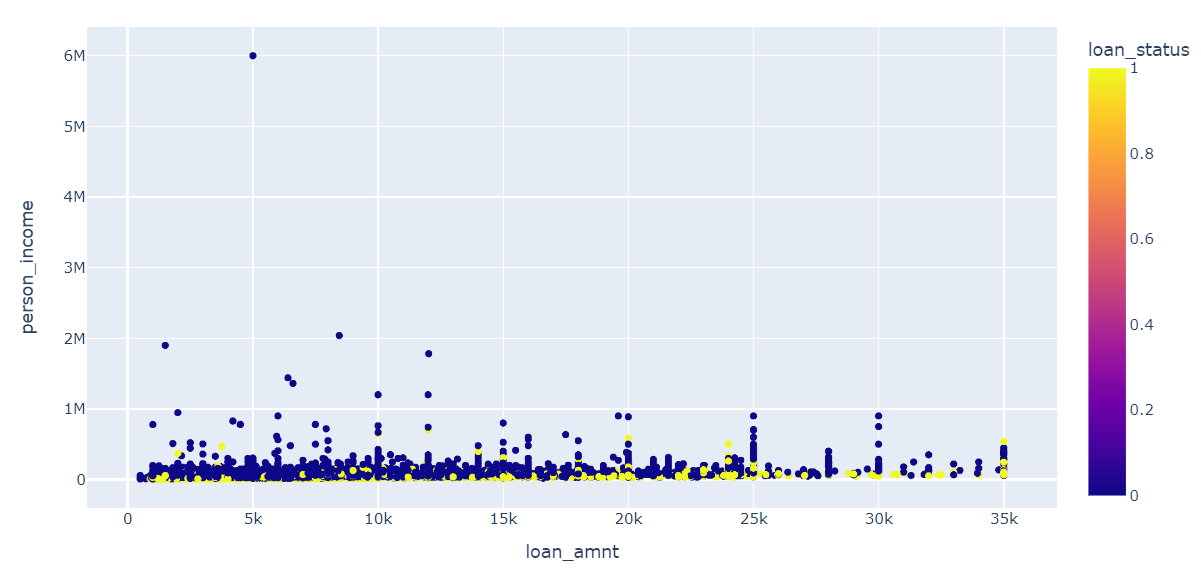
*Fig: loan interest rate vs loan grade plot*

We observed that the loan interest rate highly depends on the type of loan one is getting. So, we imputed the missing loan\_int\_rate by the median value of the interest rate of respective loan grade.

Our dataset had about 165 duplicate data which was removed. We also removed all data with person’s age more than 100. The dataset had a person with annual income of 6million, which was skewing the data. So, this data point was also removed as an outlier.

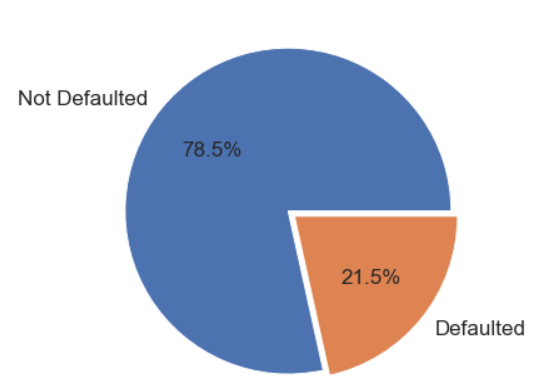
The different categorical variables were identified and changed into category datatype and then encoded. The dataset was divided into features and target and then features were scaled using sklearning standard scale and split into train test sets using train\_test\_split. Finally, the pre-processed data was fitted into different machine learning models.

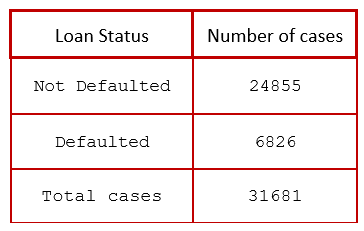
## Data visualization and analysis:

We intend to visualize and analyze the given loan default data and identify how the different variables affects the chance of default of the bank loan. For that, we firstly divided the data columns as loan details and customer personal details and then further filtered our dataset based on different bank details variables like loan amount, purpose of the loan, loan status, loan interest rate and loan grade etc. and observed their relationship with the loan default chances. We also then checked the relationship between customers personal details like their age, income, employment status, customers credit history etc. and bank default chances. We then built a prediction model based on the data and some of the results of our analysis are shown and explained below:

*Fig: Person income vs loan amount*

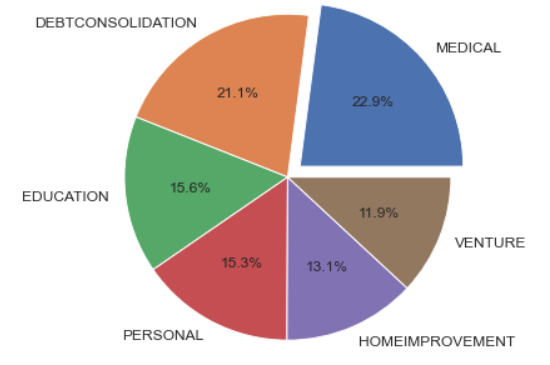
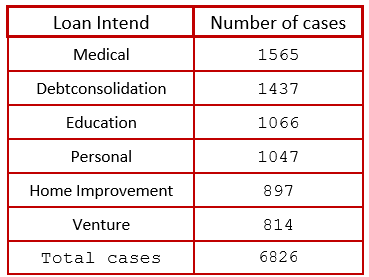
From the graph, it is observed that the chance of person taking loan is high if his/her income is low and vice versa. We observed that the annual earning for most of the people is below 2 million. Also, we observed that greater the loan amount is, greater is the chance of default as well.





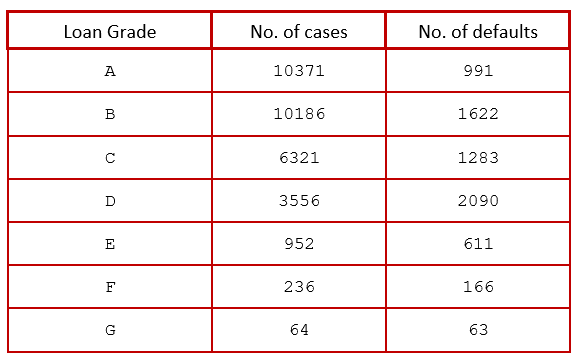
*Fig: Loan status of total issued loans*

From the dataset, it is observed that out of 31,681 issued loans, 6,826 were defaulted and 24,855 were not defaulted. The loan default percentage is 21.5% and the remaining 78.5% of loans were not defaulted.

****

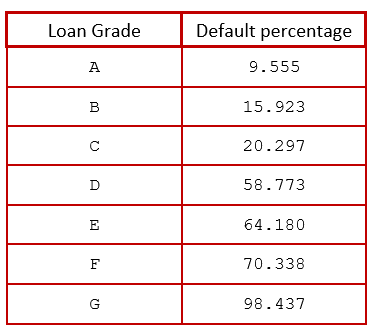
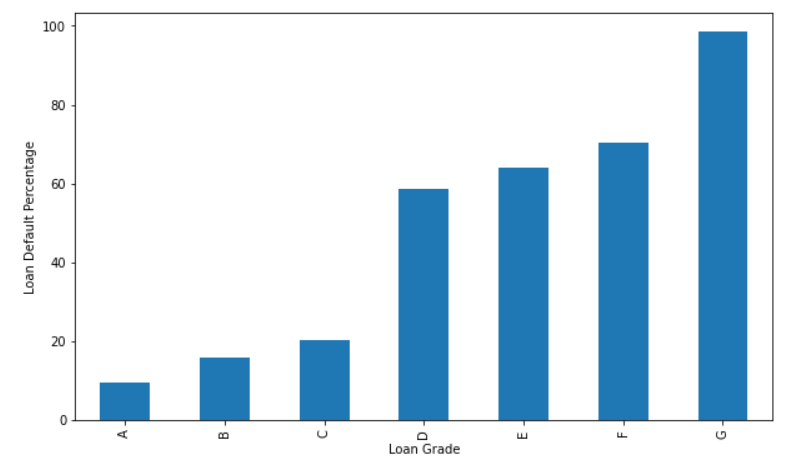
*Fig: Loan default distribution based on loan intent*

The figure illustrates the number and percentage of loan default based on different loan intents. We can observe that the loan taken for medical and debt consolidation reason got defaulted the most in comparison to others. These segments covered about 45% of the total loan default. Loan for venture was the least defaulted among all covering 11.9% of the total loan defaults.

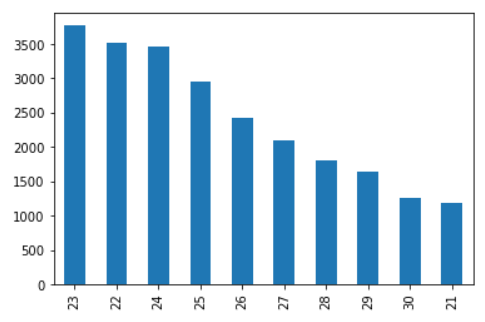
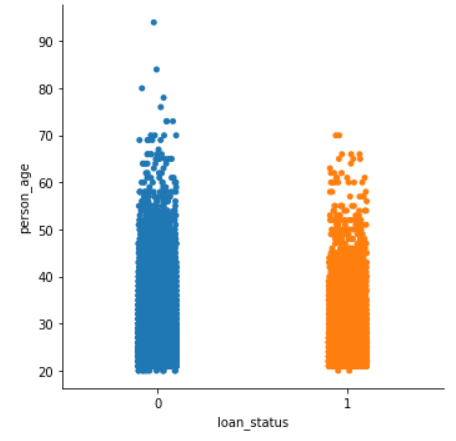


*Table: Loan default distribution based on loan grade*

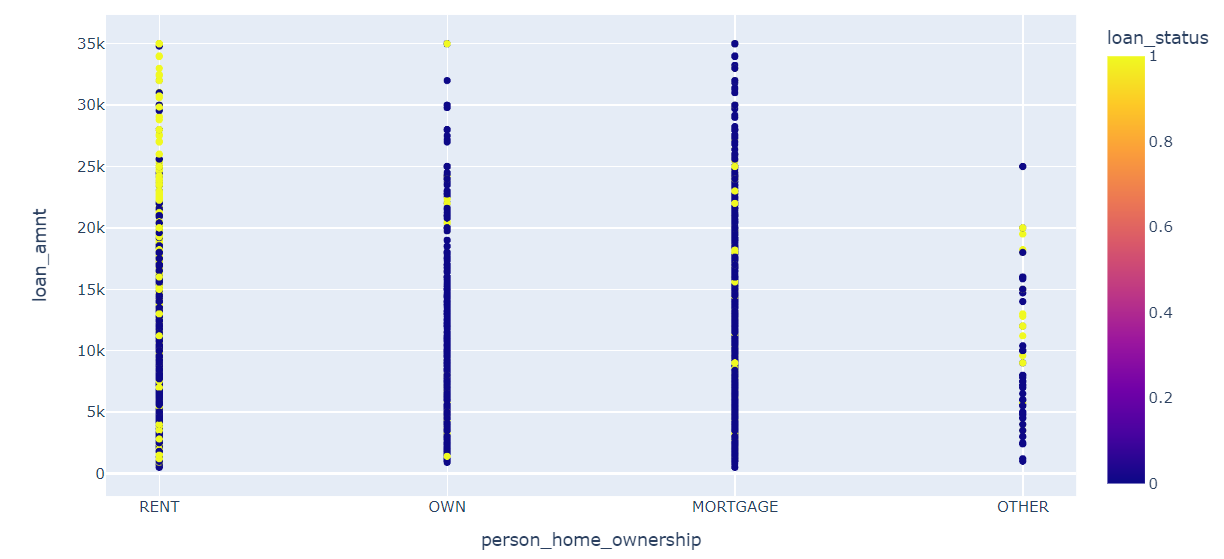
We observed that about 65% of the loans were of grade A and grade B out of the total loans that were taken from the bank. Less than 1% of the loans taken were of grade G and 63 out of 64 grade G loans taken from the bank were defaulted.



*Fig: Loan default percentage based on loan grade*

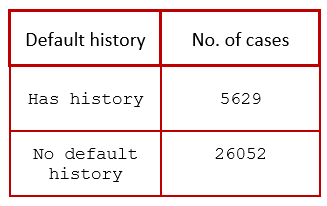
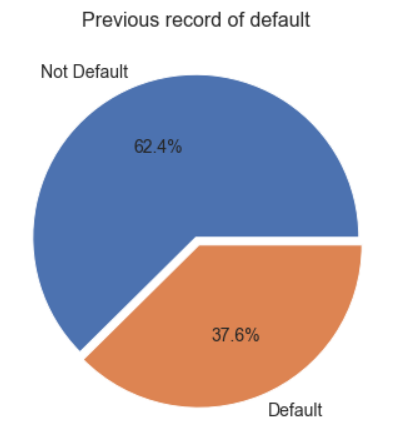
From the above figure, we can say that grade A loan has the least default chance (about 10%) in comparison to others and grade G loan has the highest chance of default (above 98%). Loans with grade D, E, F and G have more than 50 percent chance of default.

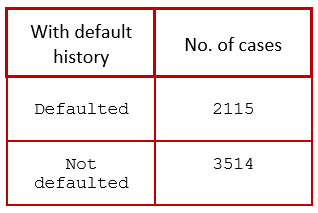
*Fig: Illustration of loan status and person age*

From the bar diagram, we observed that the person with age 23yrs takes the highest amount of loans. About 76% of the total loan were taken by the person within the age group 20 and 30. From the categorical plot, we observed that almost all borrowers default regardless of their ages.

*Fig: Illustration of loan status and person home ownership*

From the plot we can clearly see that the person who lives in rent has the highest default rate. The chance of default is very low with person who owns home completely or are paying mortgage in comparison to person in rent and others.





*Fig: Illustration of loan status and person default history*

We observed that among the 31,681 loan borrowers, 5,629 of them had previous default history i.e., about 18% of the borrowers had history record of default on the file. We also observed that about 38% of the person defaulted with pervious default history and 62.4% pay the loan amount even after having the default history in past. So, not providing loan to the person with previous default history is not appropriate.

## Predictive Modeling:

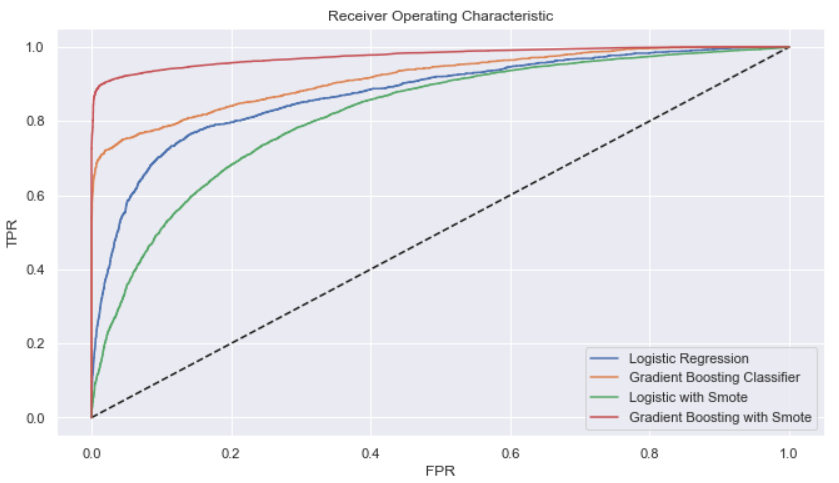
Finally, we built two different prediction models namely logistic regression model and gradient boosting classifier model with and without SMOTE (Synthetic Minority Oversampling Technique). These models can be used as a reference tool for the client and his financial institution to help make decisions on issuing loans.

The performance of these models is shown in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
| Logistic Regression | No default | 0.88 | 0.95 | 0.92 | 0.865 |
| Default | 0.77 | 0.55 | 0.64 |
| Gradient Boosting Classifier | No default | 0.92 | 0.99 | 0.95 | 0.925 |
| Default | 0.94 | 0.70 | 0.80 |
| Logistic Regression with SMOTE | No default | 0.77 | 0.70 | 0.73 | 0.743 |
| Default | 0.72 | 0.79 | 0.75 |
| Gradient Boosting Classifier with SMOTE | No default | 0.91 | 0.98 | 0.94 | 0.943 |
| Default | 0.97 | 0.91 | 0.94 |

*Fig: Table representing performance report of different models*

We observed that the logistic regression predicted the no loan default pretty well but predict the loan default with only about 64 percent F1-score. The gradient boosting classifier had about 80 percent F1-score while predicting default. Since there was class imbalance in our dataset, SMOTE was used to oversample the minority class and upon doing so we observed that the overall performance of logistic regression was degraded. On the other hand, the performance of the gradient boosting classifier increased significantly and predicted the loan default with 94 percent F1-score.



|  |  |
| --- | --- |
| **Models** | **AUC Score** |
| Logistic Regression | 0.873 |
| Gradient Boosting Classifier | 0.917 |
| Logistic Regression with SMOTE | 0.818 |
| Gradient Boosting Classifier with SMOTE | 0.976 |

*Fig: Illustration of ROC curve of different models*

The ROC curve indicated that the gradient boosting classifier with SMOTE performed better than any other models. We also observed that the AUC score for logistic regression is better without the SMOTE. Also, the highest AUC score is obtained from the gradient boosting model with SMOTE with the AUC value 0.976.

## Results:

Some of the important results and inferences generated from our analysis are enlisted below:

* 21.5% of the loans got defaulted out of 31,681 issued loans.
* G grade loans defaulted 98 percent of the time.
* A grade loans has the least chance of being default (9.55 percent).
* 62.4% of people with previous default history did not default again.
* Person who owns home has very low chance of loan default.
* About 76% of the total loan were taken by the person of age group 20–30.
* All borrowers default regardless of their ages.
* AUC score of gradient boosting with SMOTE is 0.976 which indicates that this model predicts loan default or not default correctly about 97 percent of the time.

## Conclusion and Recommendation: