Code Rush

PREPARED BY:

Pralad Kadel
Data Analyst Apprentice

Code Rush
Chakupat-10, Lalitpur
Nepal
coderush.com.np

Default Prediction Model

PROJECT REPORT

PREPARED FOR:

Shristi Shrestha
Associate Project Manager

Code Rush
Chakupat-10, Lalitpur
Nepal
coderush.com.np

1. 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.

2. Key Aspects of the Assessment

❖ Data:

The data set 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 of customer
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 on loan
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, heatmap, 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'.

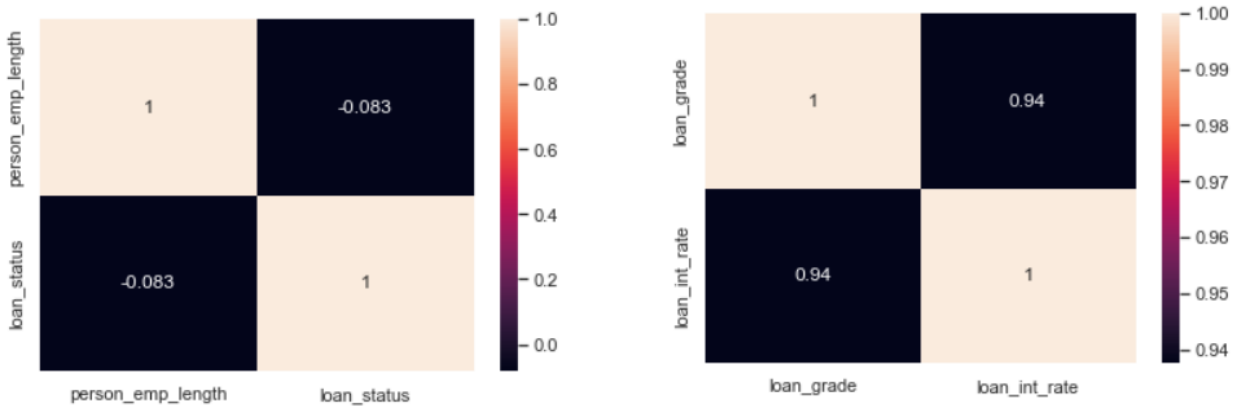


Fig: Illustration of correlation among missing values and loan status

The person employment length column had only few missing values and since it didn't have high correlation (only -0.083) with the loan default status, we dropped the missing values for this column. On the other hand, loan interest rate highly affected the default chance directly (0.94 correlation).

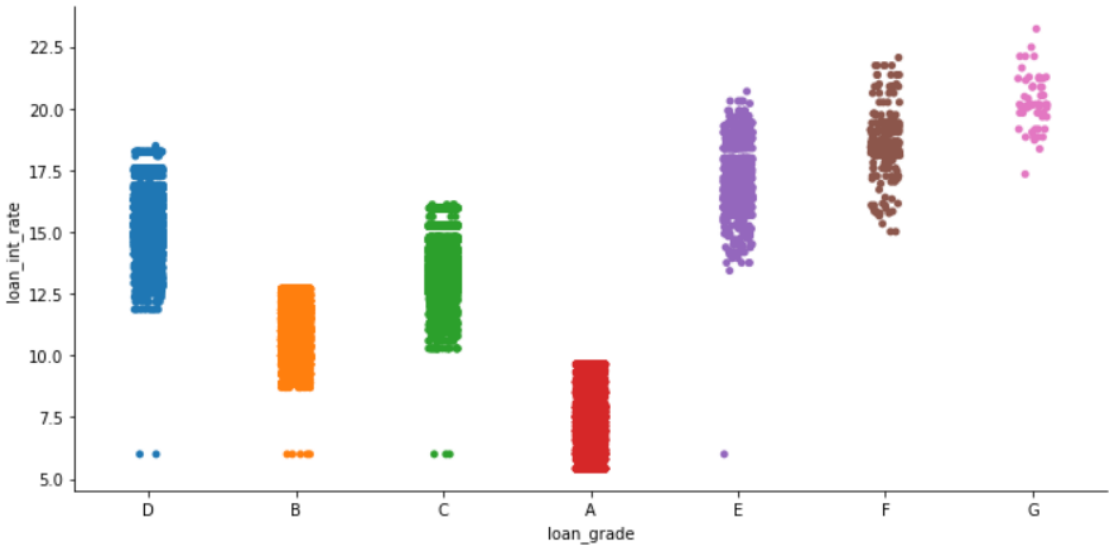


Fig: loan interest rate vs loan grade plot

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:

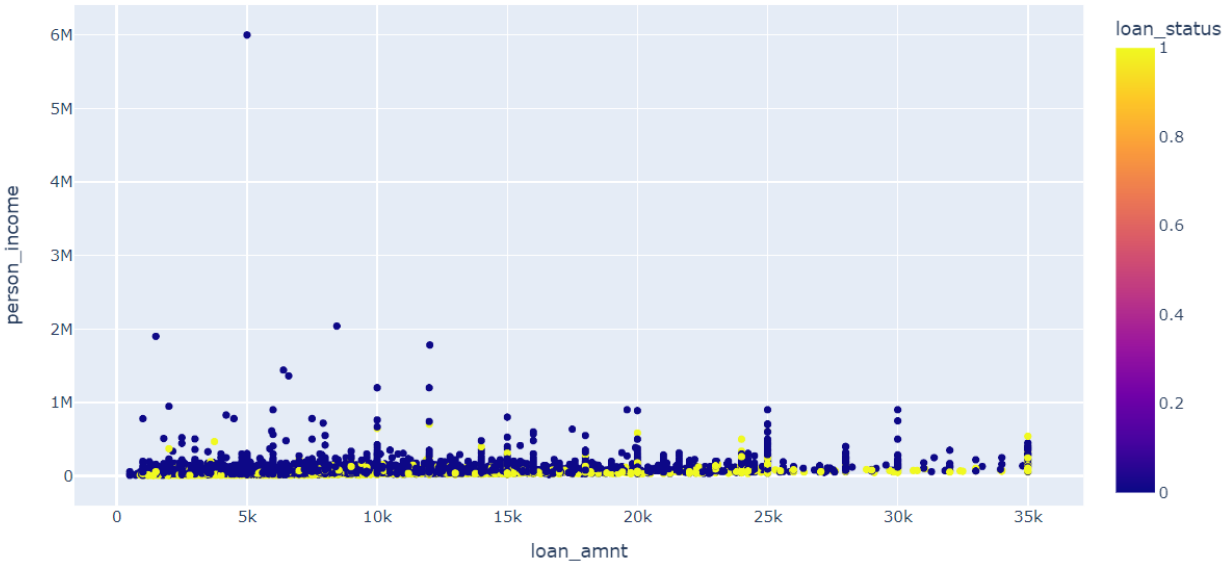


Fig: Person income vs loan amount

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.

Loan Status	Number of cases
Not Defaulted	24855
Defaulted	6826
Total cases	31681

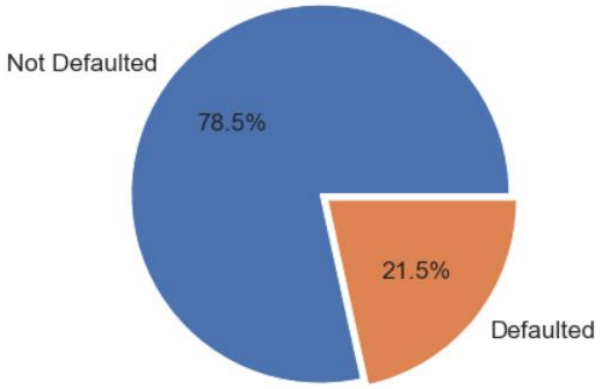


Fig: Loan status of total issued loans

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.

Loan Intend	Number of cases
Medical	1565
Debtconsolidation	1437
Education	1066
Personal	1047
Home Improvement	897
Venture	814
Total cases	6826

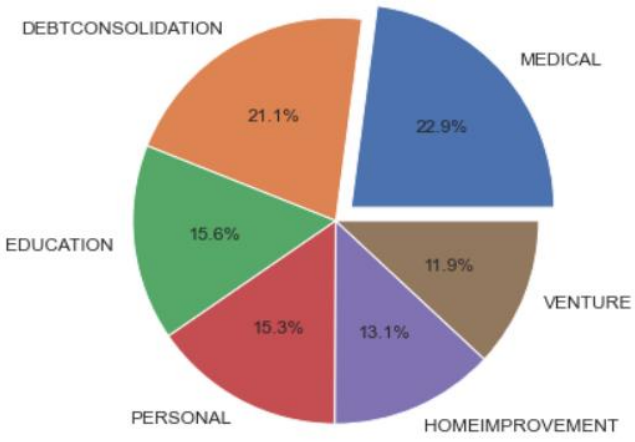


Fig: Loan default distribution based on loan intent

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.

Loan Grade	No. of cases	No. of defaults
A	10371	991
B	10186	1622
C	6321	1283
D	3556	2090
E	952	611
F	236	166
G	64	63

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.

Loan Grade	Default percentage
A	9.555
B	15.923
C	20.297
D	58.773
E	64.180
F	70.338
G	98.437

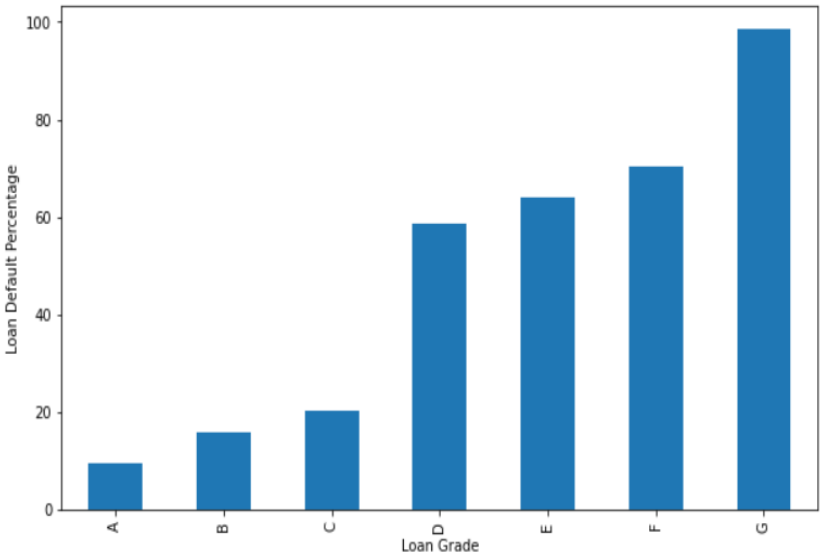


Fig: Loan default percentage based on loan grade

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.

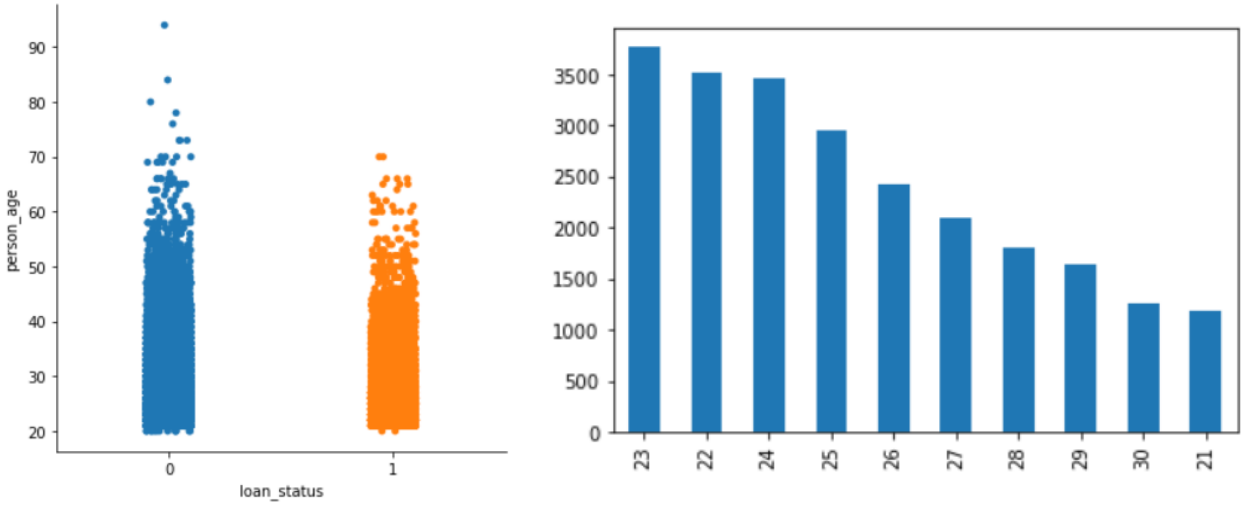


Fig: Illustration of loan status and person age

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.

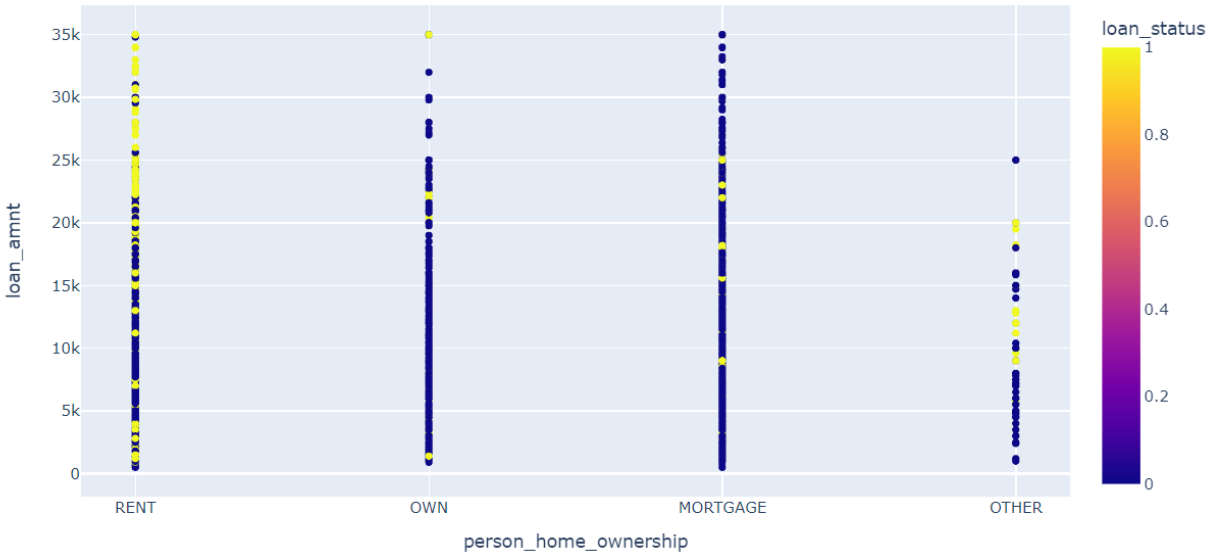


Fig: Illustration of loan status and person home ownership

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.

Default history	No. of cases
Has history	5629
No default history	26052

With default history	No. of cases
Defaulted	2115
Not defaulted	3514

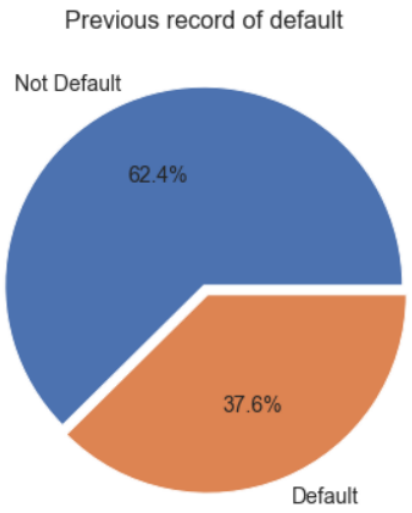


Fig: Illustration of loan status and person default history

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

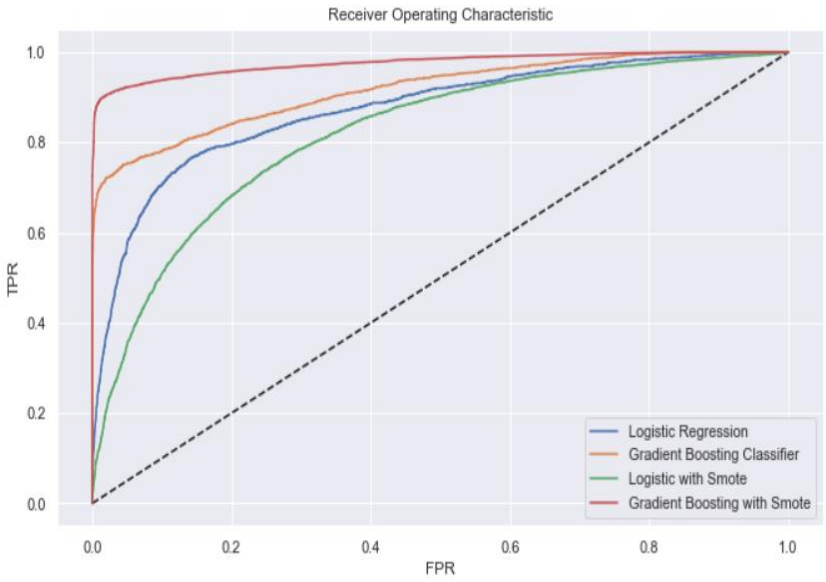


Fig: Illustration of ROC curve of different models

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 correctly predicts whether the loan defaults or not about 97 percent of the time.

❖ Conclusion and Recommendation:

From the analysis we concluded that person home ownership, loan grade, loan intent and previous default on file had the highest correlation and most impact on the overall loan default chances in comparison to other variables. Since the grade G loan had 98% chance to default, we recommend the financial institution to be very careful while issuing such loan plus increase the screening process of the candidate while issuing this grade loan.

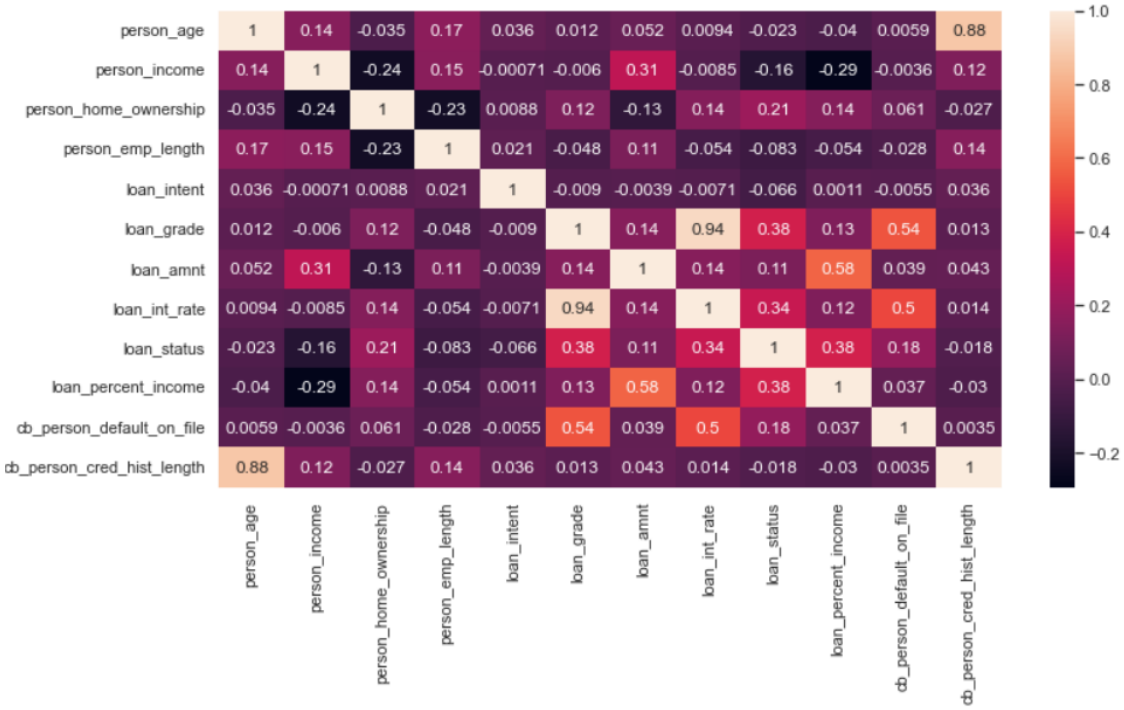
We also observed that 62.4% of people with previous default history did not default again. So, it is very unfair to not provide loan to the clients with previous default history. Thus, we recommend the financial institutions to issue loan to clients with default history by doing background check thoroughly first and then by taking reference of the machine learning models as well.

We observed that our best performing predictive model was gradient boosting model with SMOT with the AUC Score of 0.976 implying that the model correctly predicts whether the loan defaults or not about 97 percent of the time. Thus, we recommend our client to use this model as a reference tool for the loan borrowers to help make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized.

Since the dataset was taken from a single source only, we can not generalize this model for all the financial institution and banks. So, we recommend our client to not use it as a reference to predict the loan default for other institutions and banks.

❖ Appendices:

Figure representing co-relation of all features and target variables



Code to check for multi-collinearity among the feature variables

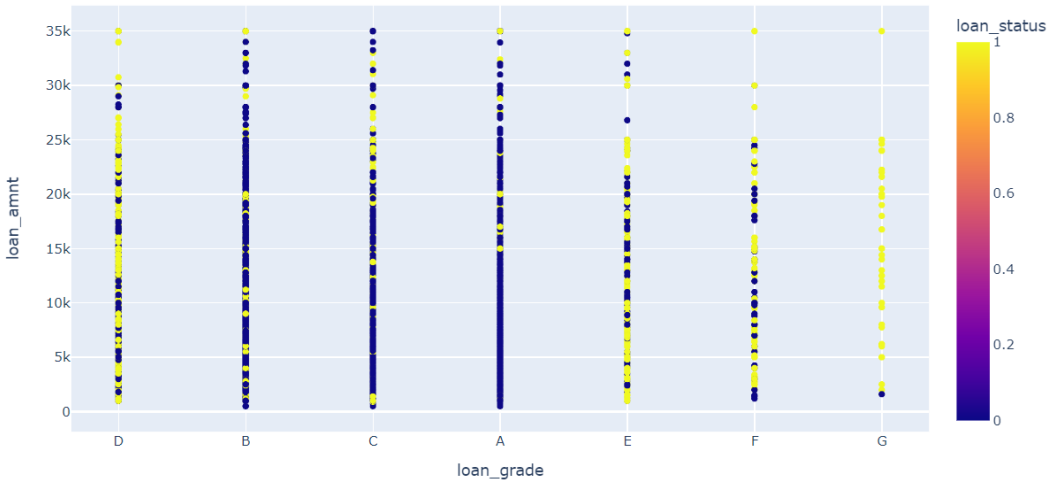
```
#check for multi-colinearity
def correlation(dataset, threshold):
    col_corr = set() # Set of all the names of deleted columns
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if (corr_matrix.iloc[i, j] >= threshold) and (corr_matrix.columns[j] not in col_corr):
                colname = corr_matrix.columns[i] # getting the name of column
                col_corr.add(colname)
                if colname in dataset.columns:
                    del dataset[colname] # deleting the column from the dataset
    print(dataset)
correlation(df,0.70)
```

Code to visualize the default rate whose loan percent income (LTI) is greater than 32%

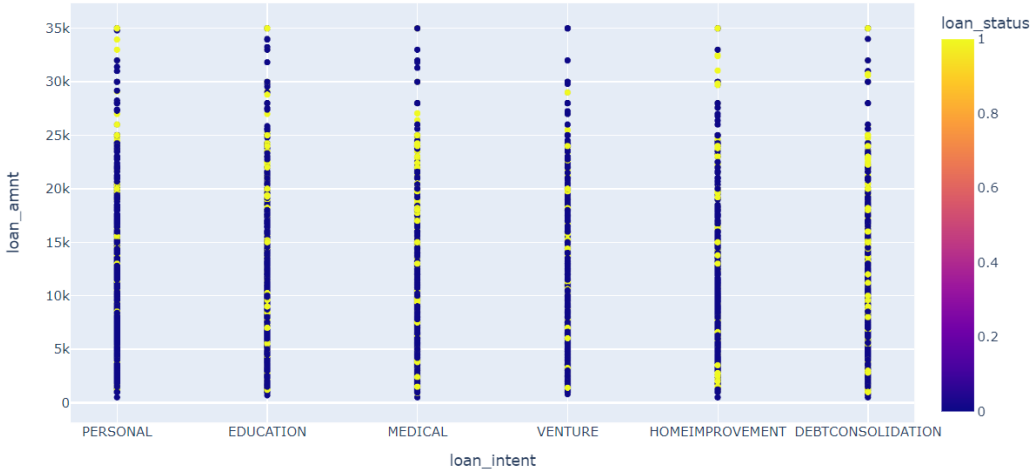
```
z = mort[mort['loan_percent_income']>0.32]
x = z['loan_status'].value_counts().sum()
print(x)
print('Default Percentage for loan with LTI greater than 32:', 201/604*100)
print('Accuracy: ', 100 - 201/604*100)
```

```
802
Default Percentage for loan with LTI greater than 32: 33.27814569536424
Accuracy:  66.72185430463577
```

Visualization of loan status based on loan amount and loan grade



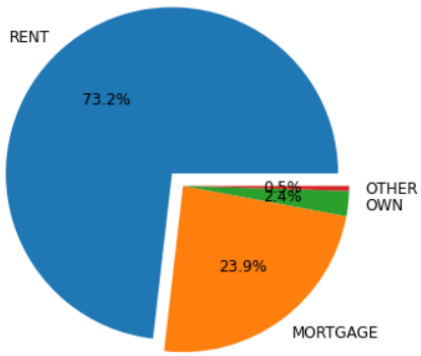
Visualization of loan status based on loan amount and intent of the loan



Visualization of loan default rate based on person home ownership

```
a.plot(kind='pie',autopct='%1.1f%%',fontsize=12,figsize=(6,6),ylabel='',explode=[0.1,0,0,0])
```

```
<AxesSubplot:>
```

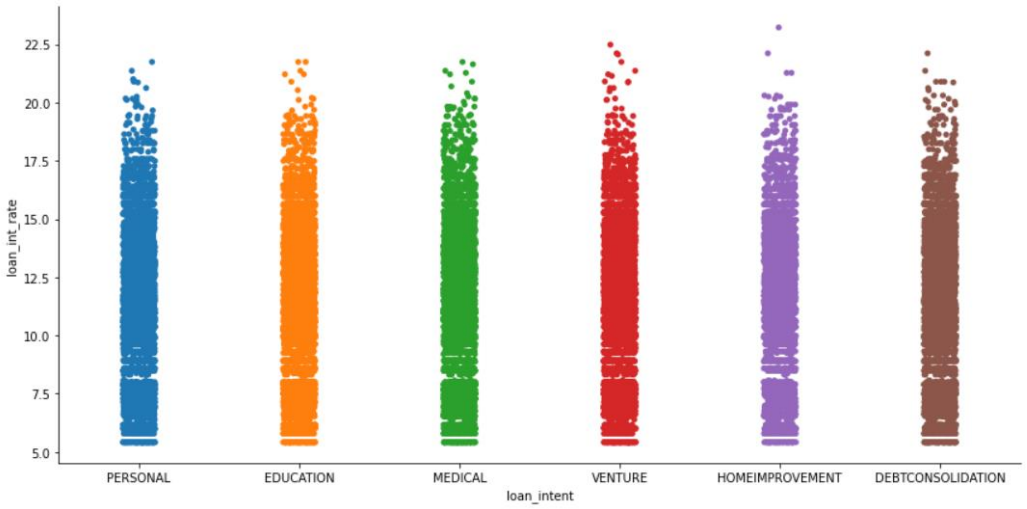


Summary Status of the continuous variables in the dataset

```
# summary stats of continuous variables
df.describe()
```

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	31529.000000	3.152900e+04	31529.00000	31529.000000	31529.000000	31529.000000	31529.000000	31529.000000
mean	27.759238	6.670447e+04	4.79051	9665.152717	11.048127	0.215928	0.169660	5.816201
std	6.366137	6.245756e+04	4.14549	6336.661842	3.203194	0.411471	0.106324	4.064782
min	20.000000	4.000000e+03	0.00000	500.000000	5.420000	0.000000	0.000000	2.000000
25%	23.000000	3.945600e+04	2.00000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.600000e+04	4.00000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	8.000000e+04	7.00000	12500.000000	13.480000	0.000000	0.230000	8.000000
max	144.000000	6.000000e+06	123.00000	35000.000000	23.220000	1.000000	0.830000	30.000000

Visualization of distribution of loan interest rate and intent of the loan



Scaling of the training dataset

```
# scale the features
from sklearn.preprocessing import StandardScaler
st = StandardScaler()
xtrain = st.fit_transform(xtrain)
xtest = st.fit_transform(xtest)
```

Encoding of the categorical variables

```
# encode the categorical variables
newdf = pd.get_dummies(df)
newdf = newdf.drop(columns='cb_person_cred_hist_length')
# newdf = newdf.drop(columns='loan_grade_G')
newdf
```

Changing the datatypes for the categorical variables

```
# change the datatypes
df["person_home_ownership"] = df["person_home_ownership"].astype("category")
df["loan_intent"] = df["loan_intent"].astype("category")
df["loan_grade"] = df["loan_grade"].astype("category")
df["cb_person_default_on_file"] = df["cb_person_default_on_file"].astype("category")
```

❖ References:

1. Reddy M. and Kavitha B., “Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis,” in Proceedings of International Conference on Signal Acquisition and Processing, Bangalore, pp. 274-277, 2010.

2. Shoumo S., Dhruba M., Hossain S., Ghani N., Arif H., and Islam H., “Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking,” in Proceedings of TENCON IEEE Region 10 Conference, Kochi, pp. 2023-2028, 2019.

3. A. Jeremy Mahoney (2020, Sep 9), Credit risk modeling with machine learning. towardsdatascience site, Last accessed 3rd Sep, 2022: Credit Risk Modeling with Machine Learning | by A. Jeremy Mahoney | Towards Data Science