Predicting Loan Defaults Using Machine Learning Model

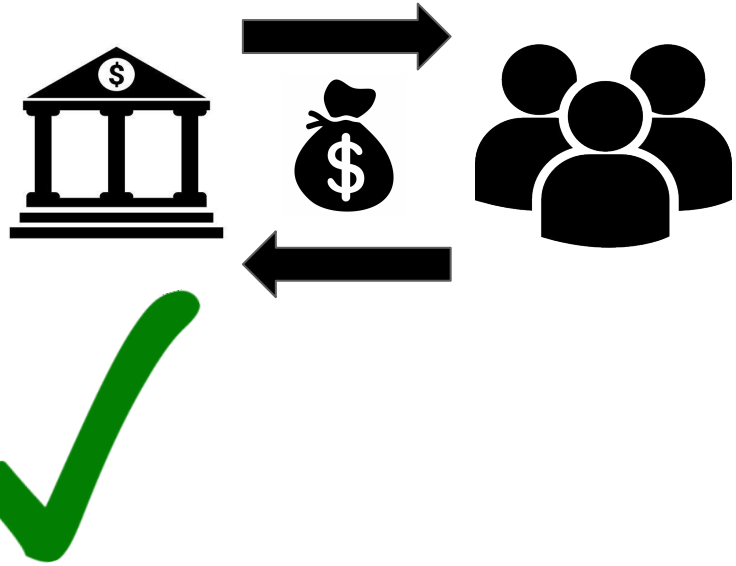
Group AS^{2}RK

Shrouq Alharbi
Sarah Alzahrani
Asmaa Alzahrani
Amani Almutairi
Rawan Alshehri
Khalid Alsuahimi

Table of Contents

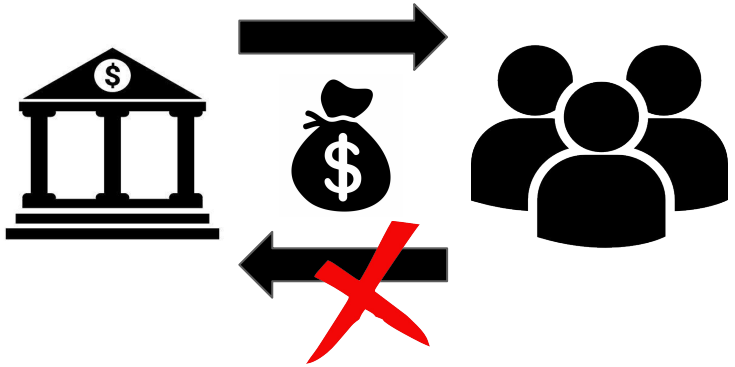
Background


JPMORGAN CHASE & CO
Bank of America
WELLS FARGO
Goldman Sachs



Problem Statement

The unfavourable behaviour when a client fails to pay a loan, resulting in business losses.



Business Solution

Building a machine learning model to comprehend the previous customers' profiles and predict future loan default risk.

Business Values

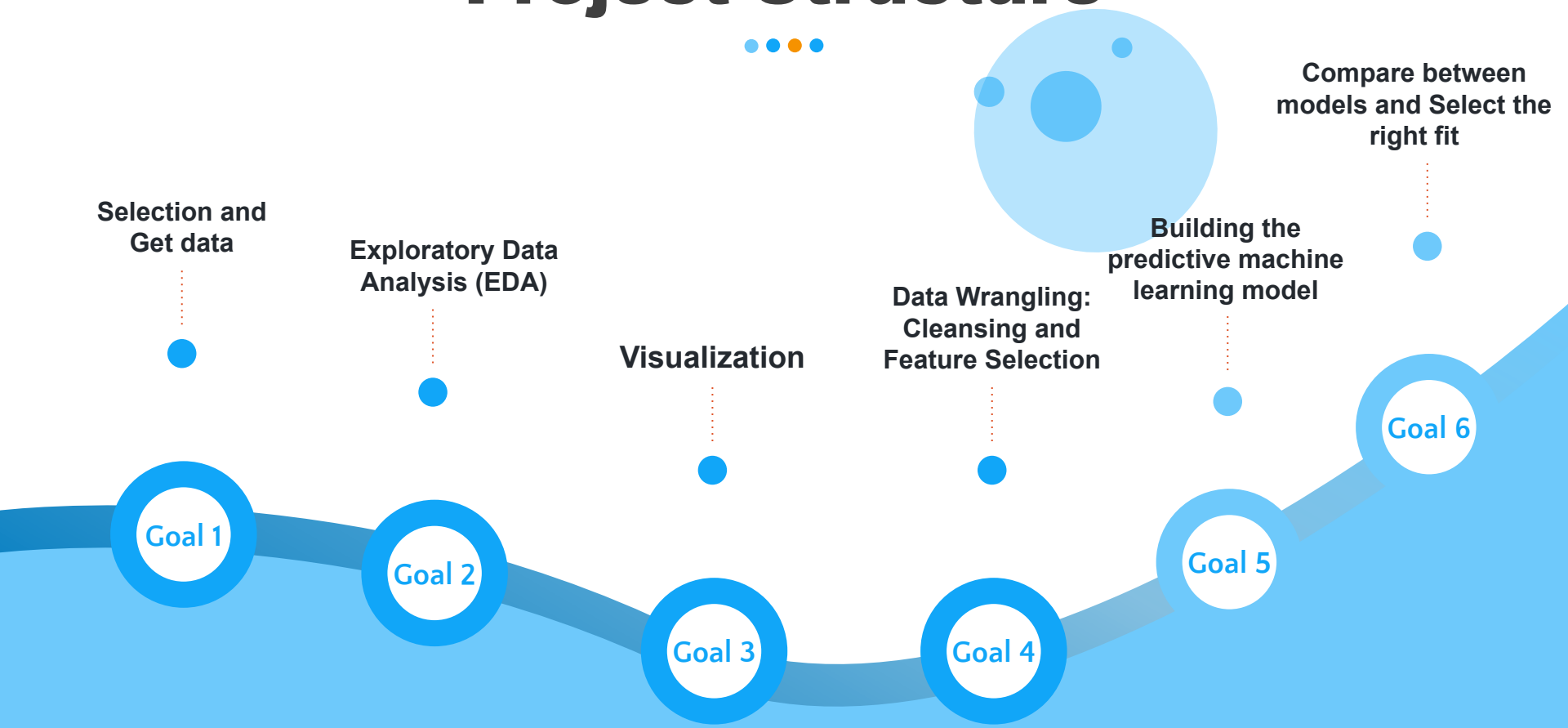
Identify the behavioral patterns that contribute to future loan default.

Assist in identifying high-risk consumers by predicting the factors impacting payback.

Enables businesses to strike a balance between risk and profit.

Implementation

Project Structure



Goal 1

Data Collection:

Selection and Get Data

Selection and Get Data

Get the Data Use pandas to read lending_club_loan.csv as a dataframe called loans

```
[13] #Loading and displaying the dataset:
     df = pd.read_csv('/content/lending_club_loan_dataset2.csv')
```

Feature description:

- **id:** Unique ID of the loan application.
- **grade:** LC assigned loan grade.
- **annual_inc:** The self-reported annual income provided by the borrower during registration.
- **short_emp:** 1 when employed for 1 year or less.
- **emp_length_num:** Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. home_ownership: Type of home ownership.
- **dti (Debt-To-Income Ratio):** A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- **purpose:** A category provided by the borrower for the loan request.
- **term:** The number of payments on the loan. Values are in months and can be either 36 or 60.
- **last_delinq_none:** 1 when the borrower had at least one event of delinquency.
- **last_major_derog_none:** 1 borrower had at least 90 days of a bad rating.
- **revol_util:** Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- **total_rec_late_fee:** Late fees received to date.
- **od_ratio:** Overdraft ratio.
- **bad_loan:** 1 when a loan was not paid.

Goal 2

Exploratory Data Analysis (EDA)

Information

+ Code + Text

```
[11] # Type of variables:
     df.dtypes.sort_values(ascending=True)
```

```
id                          int64
short_emp                   int64
emp_length_num              int64
last_delinq_none            int64
bad_loan                    int64
annual_inc                float64
dti                       float64
last_major_derog_none     float64
revol_util                float64
total_rec_late_fee        float64
od_ratio                  float64
grade                      object
home_ownership             object
purpose                    object
term                       object
dtype: object
```

```
# Couting variables by type:
df.dtypes.value_counts()
```

```
float64    6
int64      5
object     4
dtype: int64
```

```
<bound method DataFrame.info of
0      11454641     A    100000.0  ...
1       9604874     A     83000.0  ...
2       9684700     D     78000.0  ...
3       9695736     D     37536.0  ...
4       9795013     D     65000.0  ...
...         ...   ...         ...  ...
19995   6595657     B     27000.0  ...
19996   1576331     B     45000.0  ...
19997   6645736     B    104000.0  ...
19998   6625736     A     38400.0  ...
19999   6625685     B    150000.0  ...

[20000 rows x 15 columns]>
```

Description

```
#Main stats of numeric attributes:
df.describe()
```

	id	annual_inc	short_emp	emp_length_num	dti	last_delinq_none	last_major_derog_none	revol_util
count	2.000000e+04	20000.000000	20000.000000	20000.00000	19846.000000	20000.000000	574.000000	20000.000000
mean	7.590662e+06	73349.578350	0.112500	6.82140	16.587841	0.546600	0.759582	55.958148
std	1.609593e+06	45198.567255	0.315989	3.77423	7.585812	0.497836	0.427710	42.117456
min	5.860400e+05	8412.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
25%	6.206283e+06	47000.000000	0.000000	3.00000	10.852500	0.000000	1.000000	38.800000
50%	7.378896e+06	65000.000000	0.000000	7.00000	16.190000	1.000000	1.000000	57.100000
75%	8.766235e+06	88000.000000	0.000000	11.00000	22.060000	1.000000	1.000000	73.900000
max	1.145464e+07	1000000.000000	1.000000	11.00000	34.990000	1.000000	1.000000	5010.000000

Checking for Missing Values

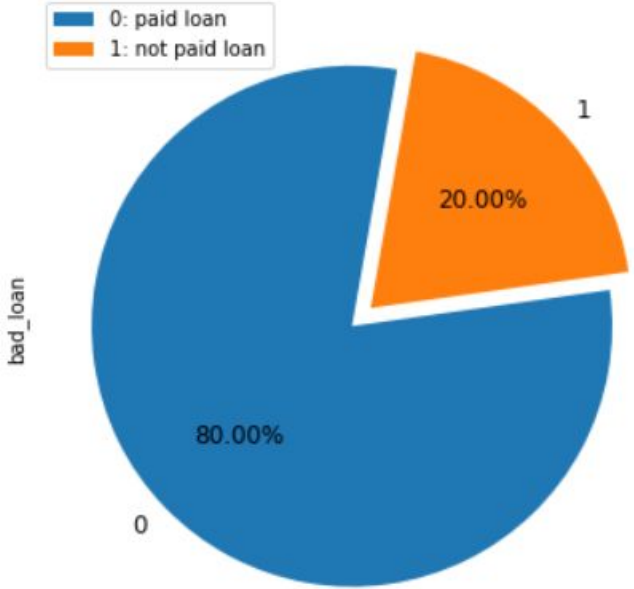
	absolute	percent %
id	0	0.00
grade	0	0.00
annual_inc	0	0.00
short_emp	0	0.00
emp_length_num	0	0.00
home_ownership	1491	7.08
dti	154	0.73
purpose	0	0.00
term	0	0.00
last_delinq_none	0	0.00
last_major_derog_none	19426	92.19
revol_util	0	0.00
total_rec_late_fee	0	0.00
od_ratio	0	0.00
bad_loan	0	0.00

Goal 3

Visualization

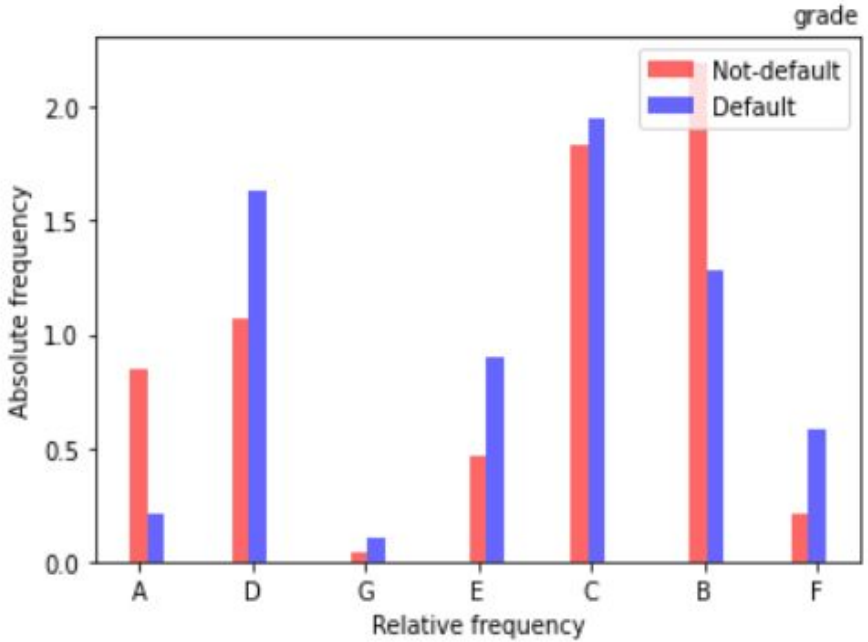
Visualization cont.

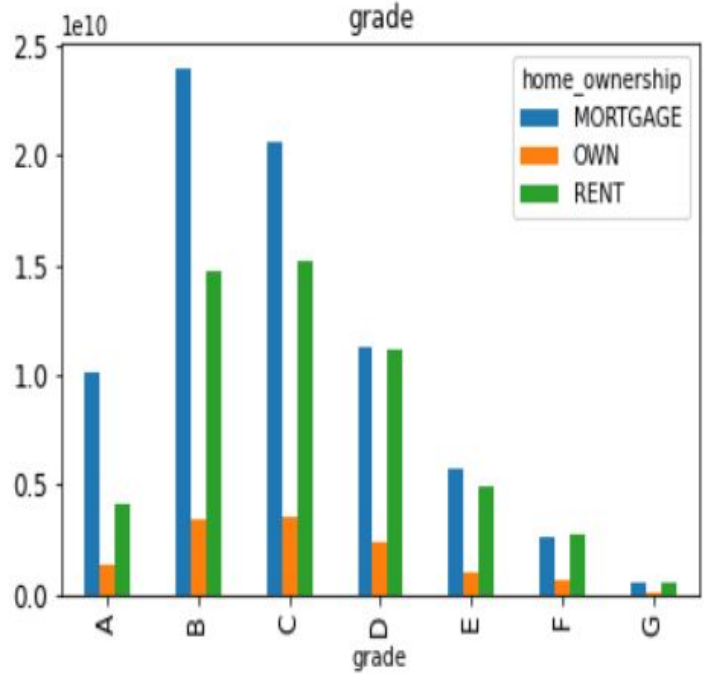
The target variable is: bad_loan 1, (default) not paid 0 paid



Visualization cont.

Feature: grade

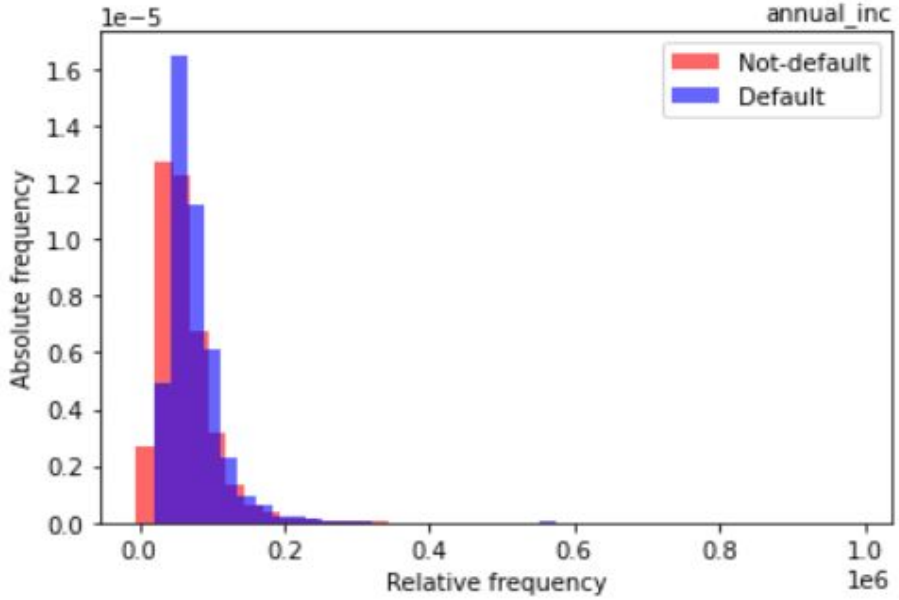




Visualization cont.

Feature: annual_inc

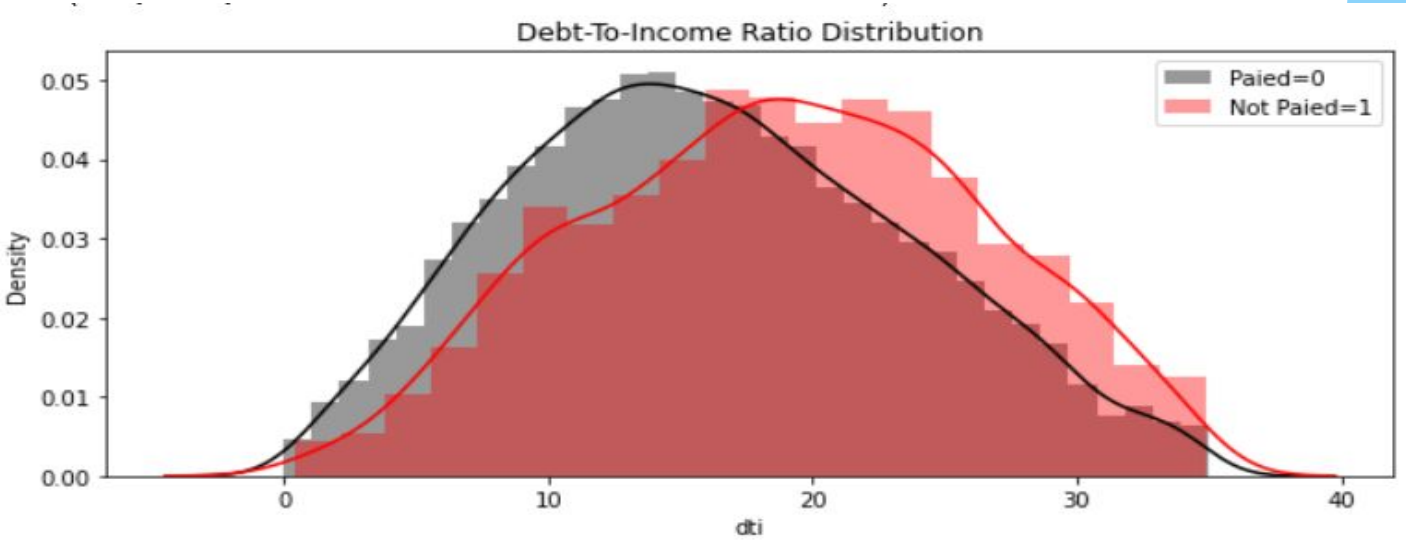
The self-reported annual income provided by the borrower during registration.



Visualization cont.

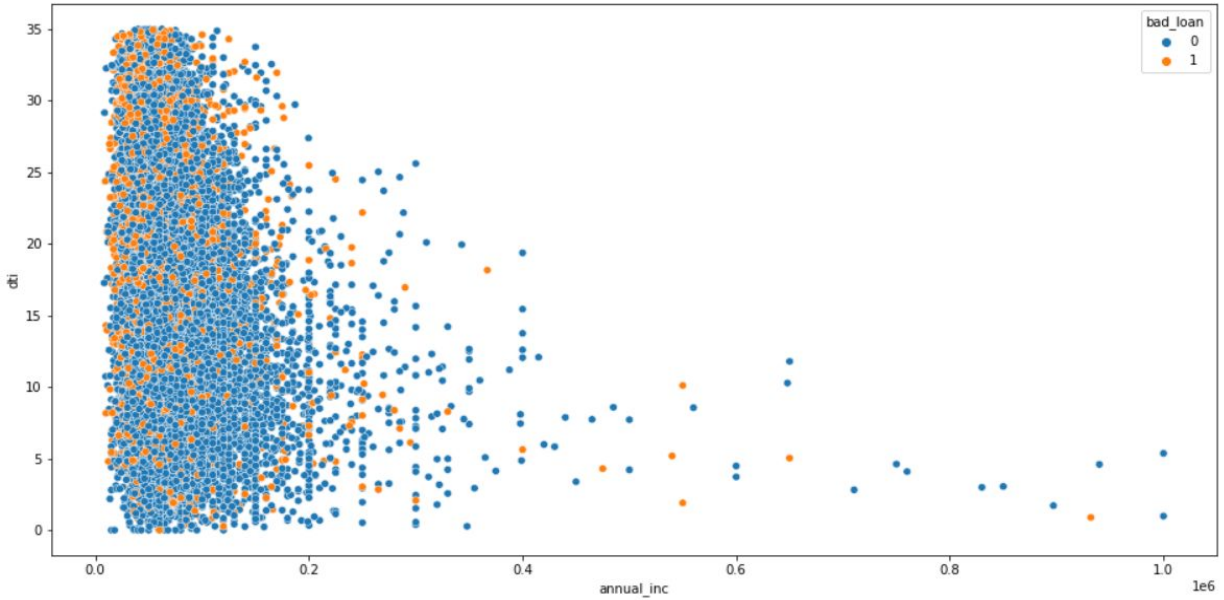
Feature: dti (debt-to-income rate)

A ratio calculated using the borrower's total monthly debt payments on the total debt obligations divided by the borrower's self-reported monthly income.



Visualization cont.

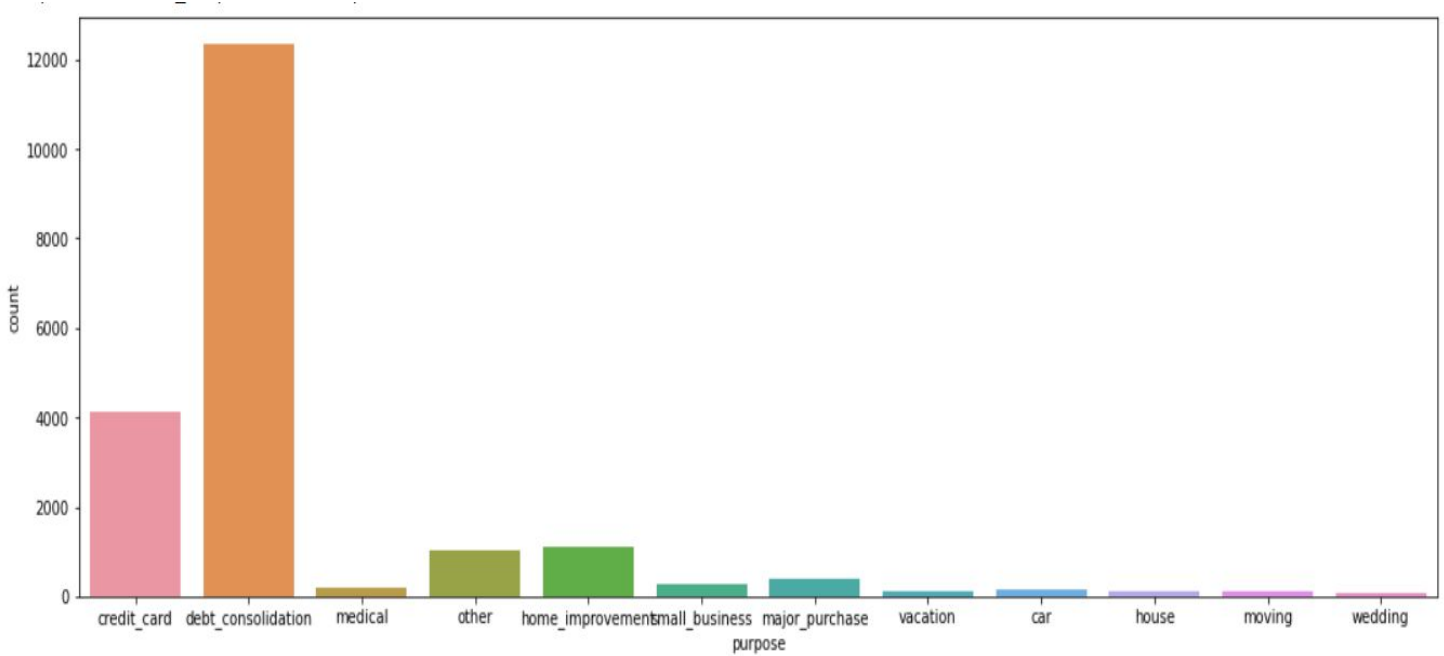
Display the distribution of ('dti', 'annual_inc', 'bad_loan')



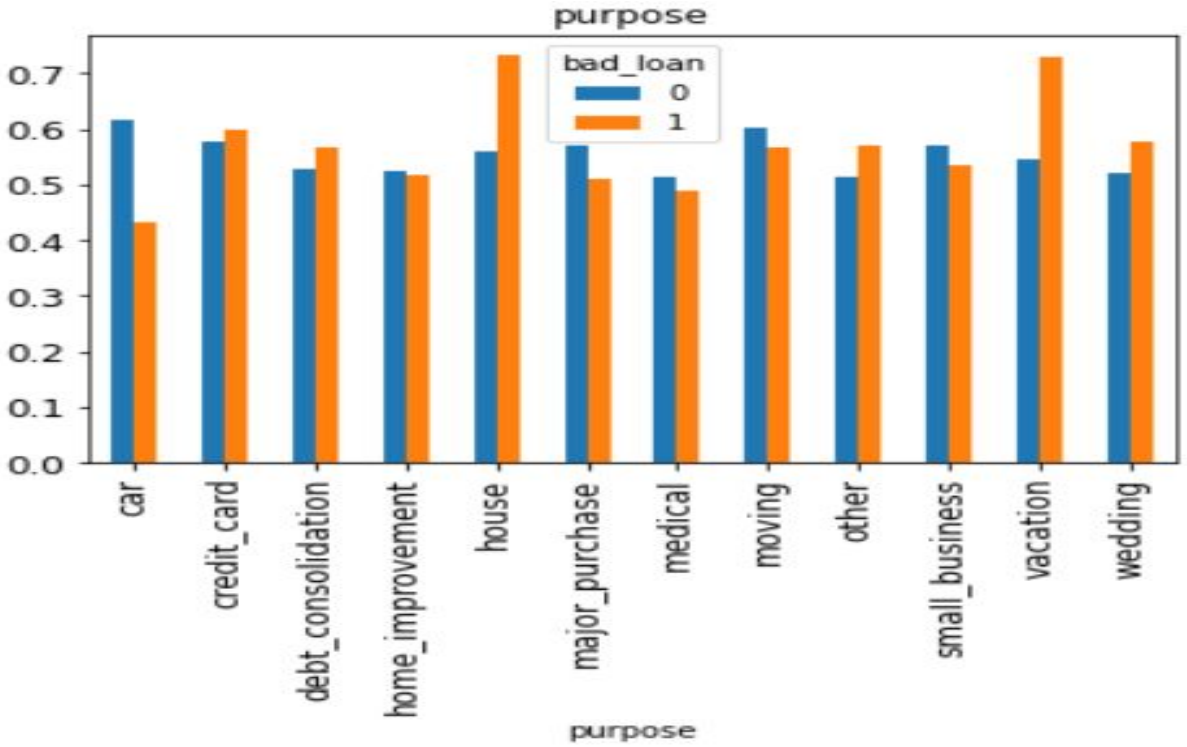
Visualization cont.

Feature: purpose

A category provided by the borrower for the loan request.



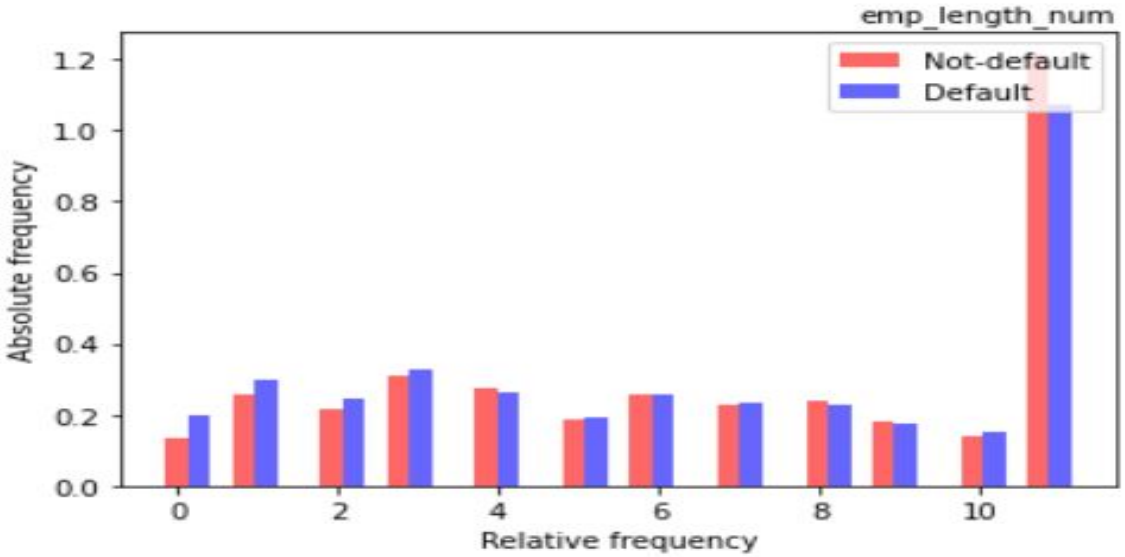
Visualization cont.



Visualization cont.

Feature: emp_length_num

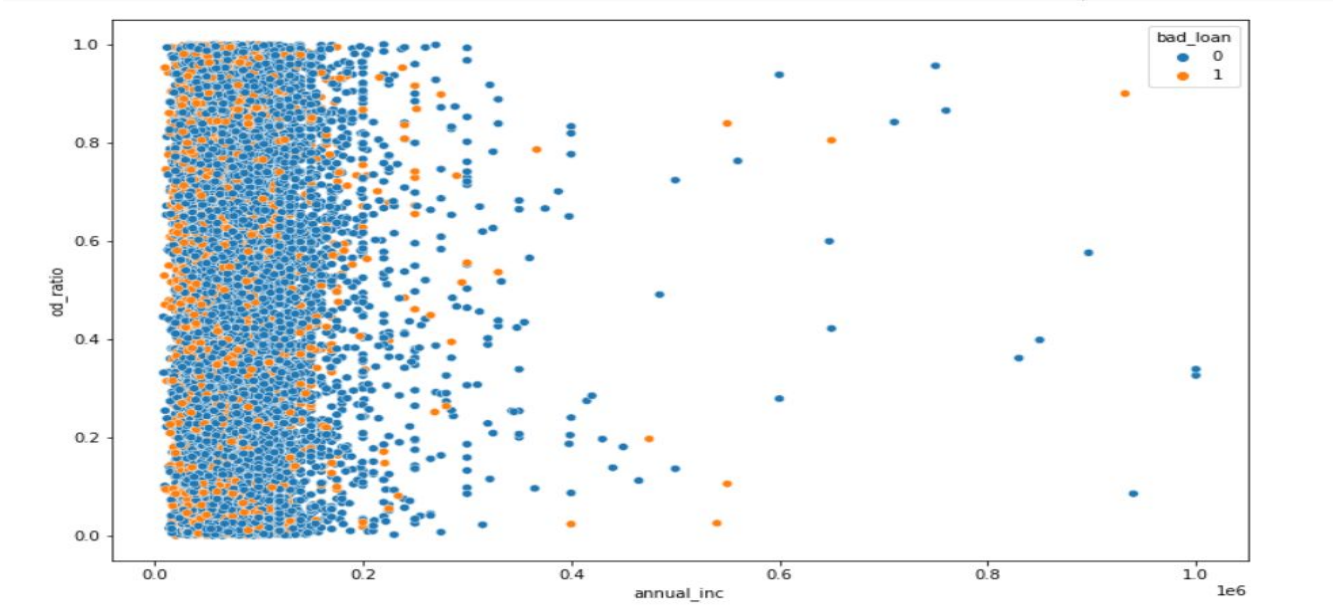
Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.



Visualization cont.

Feature: od_ratio (overdraft rate)

The scatter plot displays the correlations between 'annual_inc', 'od_ratio', 'bad_loan'

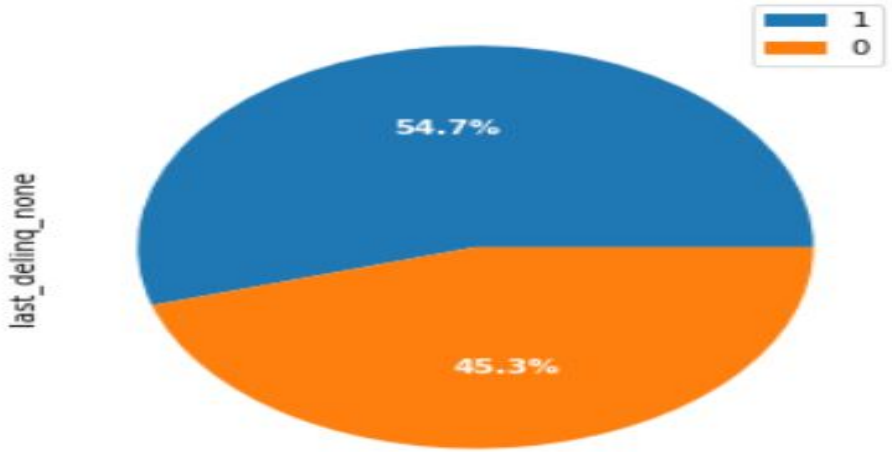


Visualization cont.

Feature: last_delinq_none

when the borrower had at least one event of delinquency

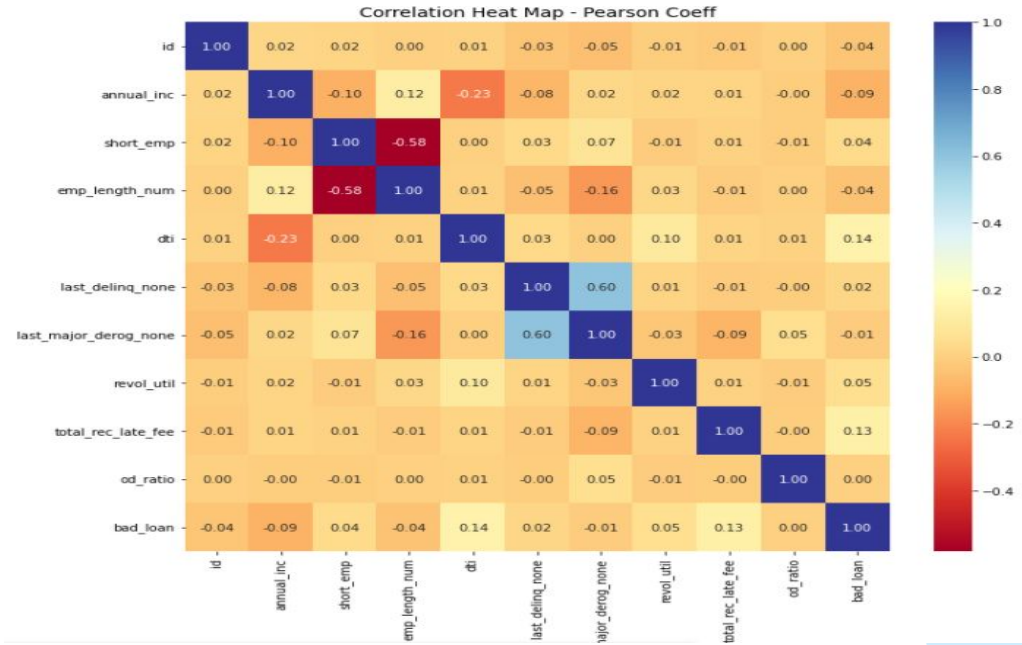
1	10932	54.66
0	9068	45.34



Visualization cont.

Heatmap

The heatmap shows there are some positive and negative correlations amongst variables.

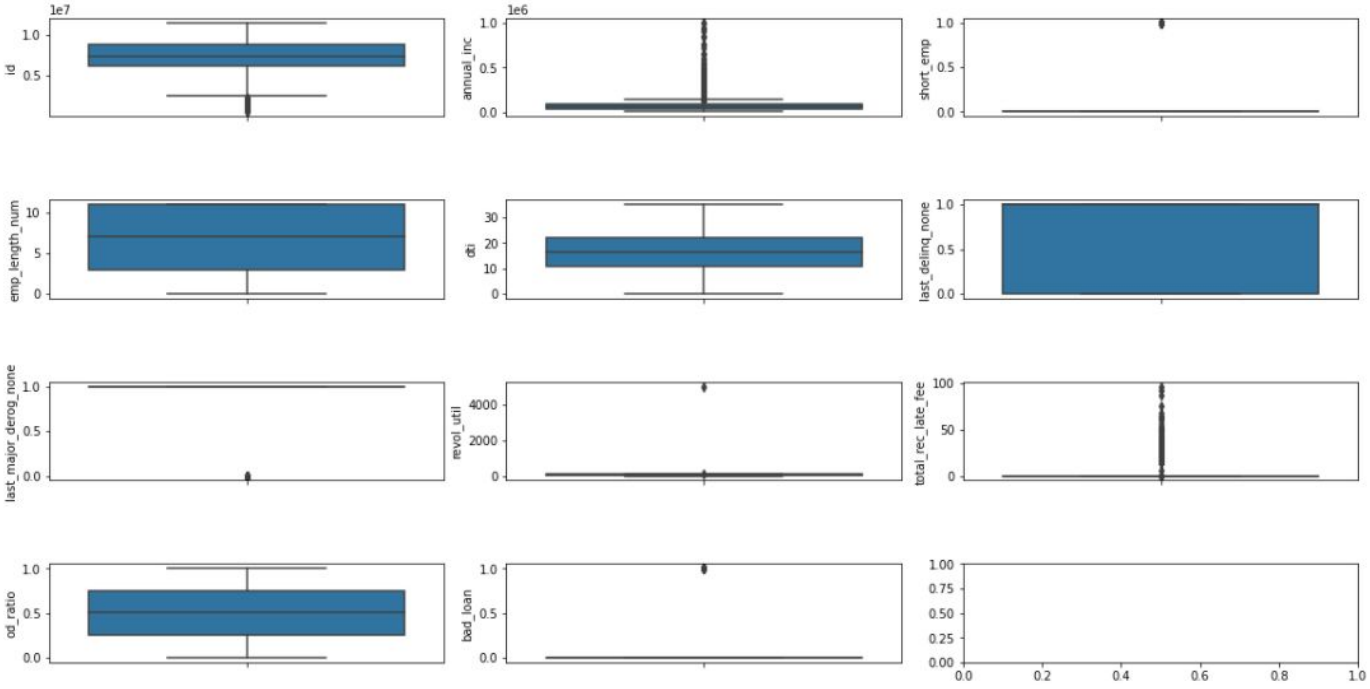


Goal 4

Data Wrangling: Cleansing and Feature Selection

Data Wrangling

Visualizing the numeric data dispersion

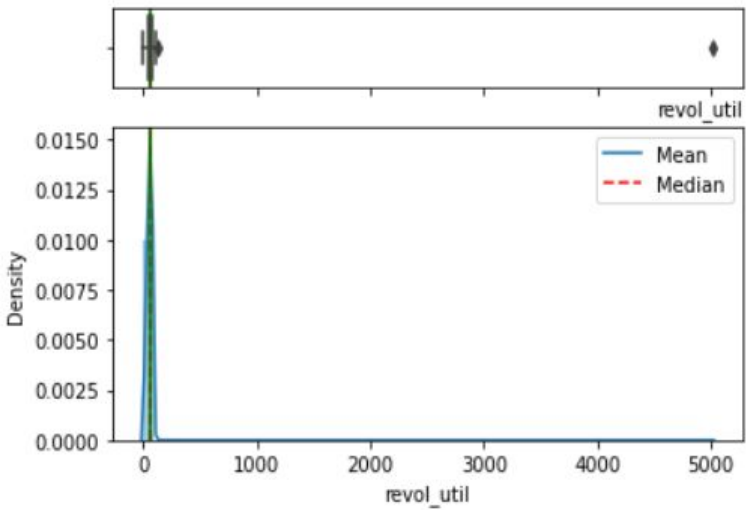


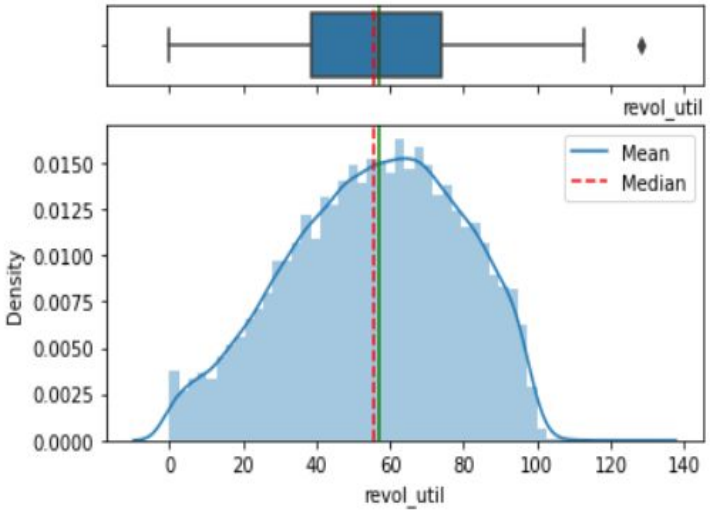
Data Wrangling cont.

Dealing with outliers:

There are some outliers in the variables ‘annual_inc’, ‘revol_util’ and ‘total_rec_late_fee’.

for example ‘revol_util’:





Data Wrangling cont.

Dealing with missing values

```
#Time to detect and eliminate them.
for column in data.columns:
    if data[column].isna().sum() != 0:
        missing = data[column].isna().sum()
        portion = (missing / data.shape[0]) * 100
        print(f"'{column}': number of missing values '{missing}' ---> '{portion:.3f}%'")
```

```
'annual_inc': number of missing values '915' ---> '4.626%'
'home_ownership': number of missing values '1476' ---> '7.462%'
'dti': number of missing values '152' ---> '0.768%'
'last_major_derog_none': number of missing values '19208' ---> '97.113%'
```

Data Wrangling cont.

Dealing with missing values

Column	Strategy
‘annual_inc’	Replacing missing values with the mean (average)
‘home_ownership’	Mode imputation (replacing NaN by most frequent value: Mortage).
‘dti’	Replacing missing values with the mean (average).
‘last_major_derog_none’	Drop ‘last_major_derog_none’ numerical variable (too many anomalies).

Goal 5

Building the Predictive Machine Learning Model

Building the Predictive ML Model

ROC Area under the Curve (AUC)

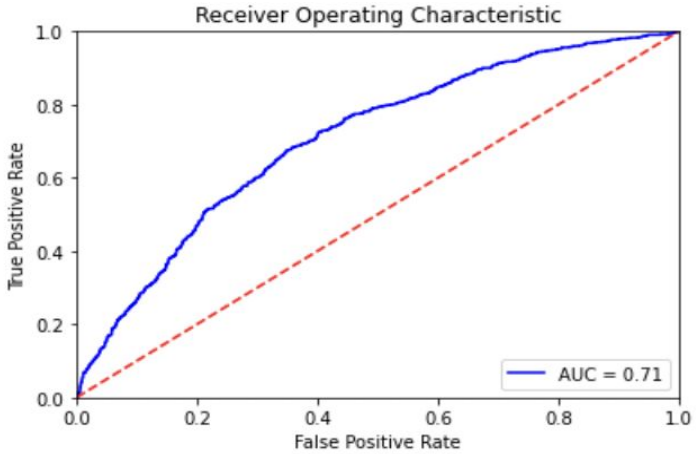
```
# ROC Curve: Area Under the Curve

def auc_roc_plot(y_test, y_preds):
    fpr, tpr, thresholds = roc_curve(y_test,y_preds)
    roc_auc = auc(fpr, tpr)

    print(roc_auc)


    plt.title('Receiver Operating Characteristic')
    plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
    plt.legend(loc = 'lower right')
    plt.plot([0, 1], [0, 1],'r--')
    plt.xlim([0, 1])
    plt.ylim([0, 1])
    plt.ylabel('True Positive Rate')
    plt.xlabel('False Positive Rate')
    plt.show()
```

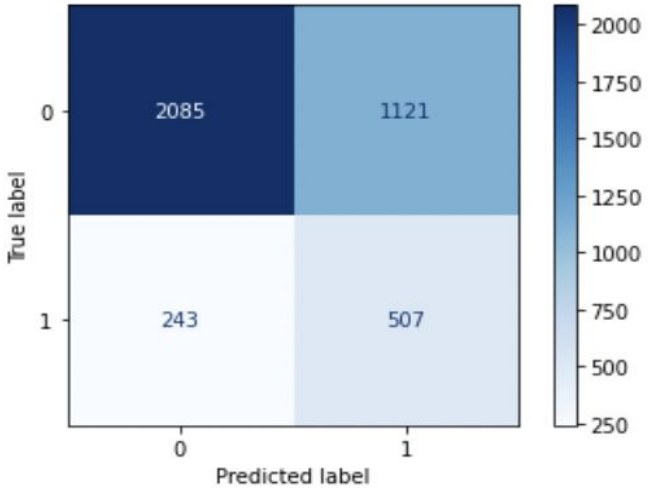
Building the Predictive ML Model cont.

Logistic Regression (LR)

```
[ ]  auc_roc_plot(y_test_lr, y_preds_lr)
```

```
0.7066791432730296
```



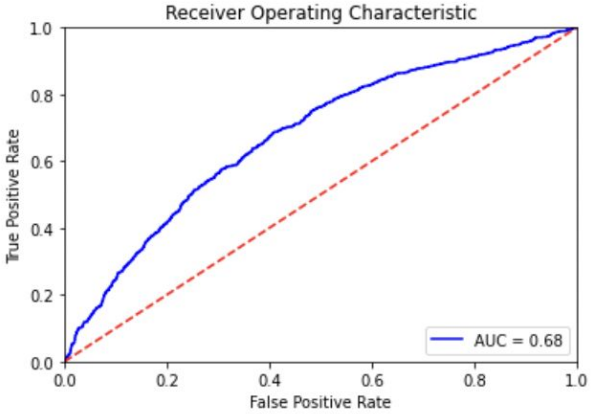


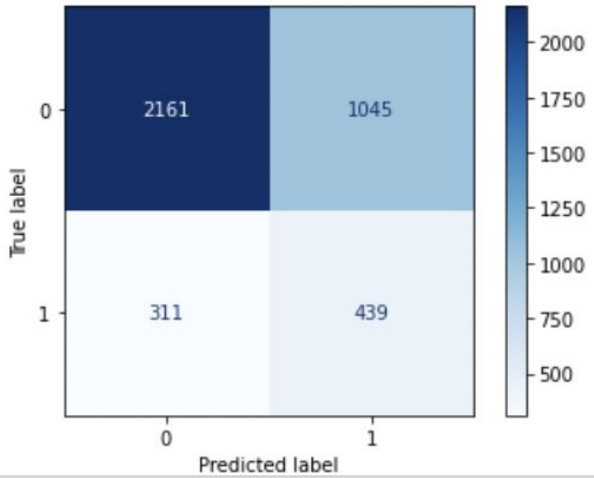
Building the Predictive ML Model cont.

Support Vector Machine (SVM)

```
auc_roc_plot(y_test_svc, y_preds_svc)
```

```
0.6762416302765648
```



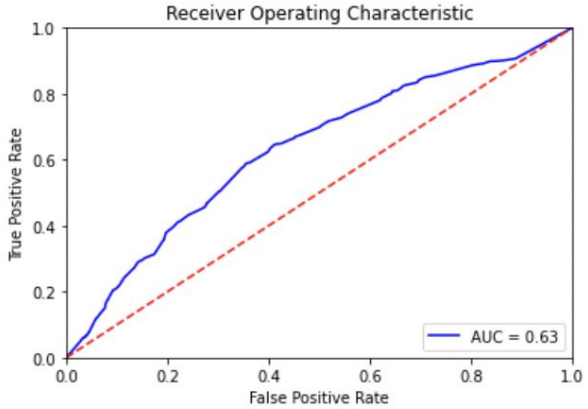


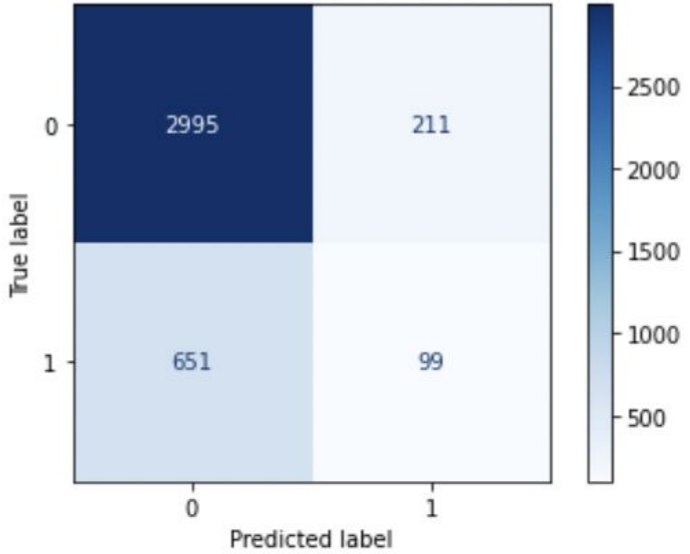
Building the Predictive ML Model cont.

Decision Trees (DT)

```
y_preds_dt = clf_tree.predict_proba(X_test_dt)[:,1]
auc_roc_plot(y_test_dt, y_preds_dt)
```

```
0.6333266791432729
```



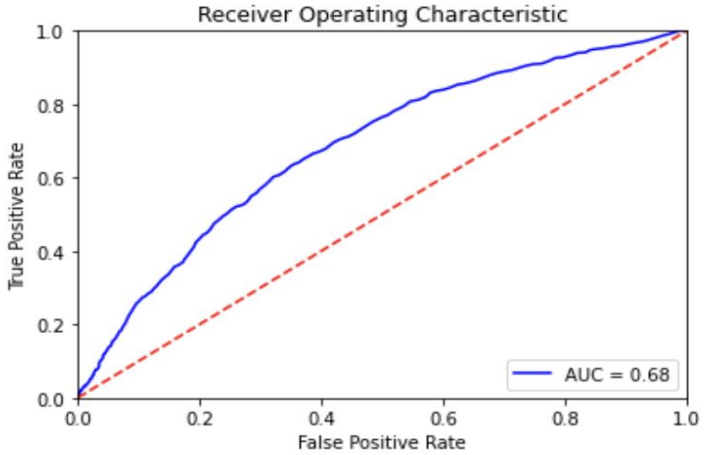


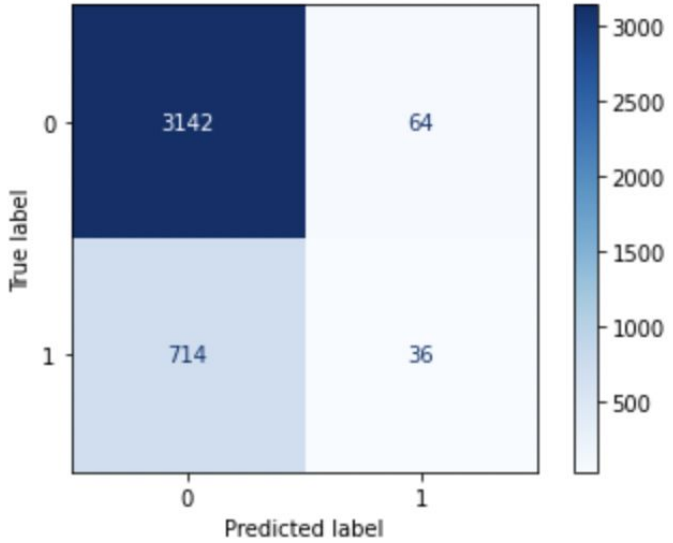
Building the Predictive ML Model cont.

Random Forest (RF)

```
auc_roc_plot(y_test_rf, y_preds_rf)
```

```
0.6831790393013101
```





Goal 6

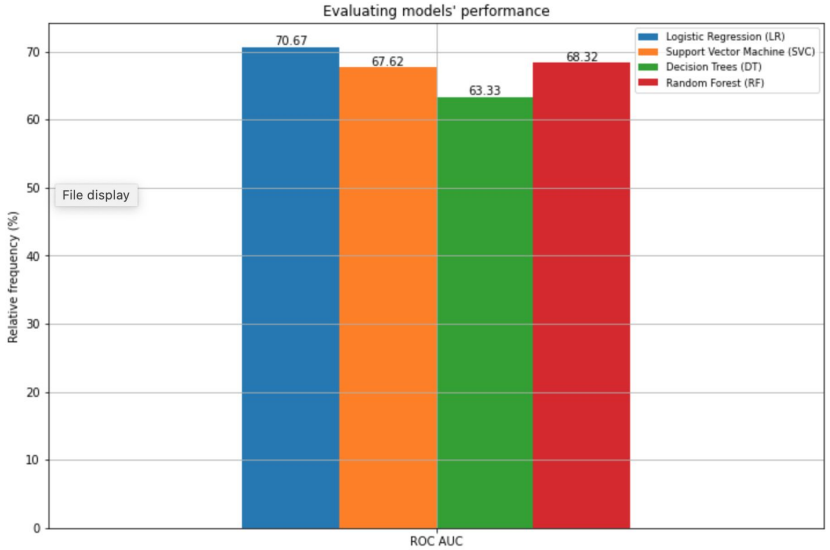
Compare between Models and Select the Right Fit

Comparison

Best model:

Logistic Regression - Classifier (LR): 70.67%.

Classifier	AUC ROC (%)
Logistic Regression (LR)	70.667914
Support Vector Machine (SVC)	67.624163
Decision Trees (DT)	63.332668
Random Forest (RF)	68.317904



Conclusion

Project Limitation

- Other machine learning models can be explored to discover the best fit.

- The model may be validated using local and Saudi datasets.

Conclusion

Through using ML algorithms:

- Avoid lending to high-risk consumers and closely monitor loan repayments.
- Avoid bad loans by ensuring that loans are issued to borrowers who are likely to repay.
- Help to conduct credit analysis of potential borrowers to determine the borrower's credit risk.

References

1.Gupta, A. (2021). Steps to Complete a Machine Learning Project - Analytics Vidhya. Retrieved 5 January 2022, from https://www.analyticsvidhya.com/blog/2021/04/steps-to-complete-a-machine-learning-project/

2.🏦 Lending Club Loan 💰 Defaulters 🏃 Prediction. (2021). Retrieved 5 January 2022, from https://www.kaggle.com/faressayah/lending-club-loan-defaulters-prediction

3.Newest 'python' Questions. Retrieved 6 January 2022, from https://stackoverflow.com/questions/tagged/python

4.Xu, Z. (2021). Loan Default Prediction for Profit Maximization. Retrieved 5 January 2022, from https://towardsdatascience.com/loan-default-prediction-for-profit-maximization-45fcd461582b

5.Kolawole, P. (2020). Can you Predict Customer's Loan Default using Machine Learning?. Retrieved 5 January 2022, from https://medium.datadriveninvestor.com/can-you-predict-customers-loan-default-using-machine-learning-be774489b8f5

THANK YOU!

Any Questions?