Problem Statement

Background: A well-established bank in Asia Pacific is looking to use data to better their nonperforming loan ratio.

The bank aims to predict whether a loan will be paid back to the bank or not. To solve this problem, the bank needs to gather the right data, analyse it and create a model to predict whether a loan will be paid back or not. Since this is the first time the bank is attempting to use data actively, they will need help on finding the right data to predict nonperforming loans and if a client is able to pay the due instalment. At the moment, the bank is fully dependent on the credit risk rating and their past experience. However, the bank is unable to accurately predict whether a client is able to pay their loan back on a continuous basis, month to month, and hence is unable to serve their clients better and prevent clients from defaulting.

The banks goals are to:

predict if a client is able to pay the instalment in any given month predict how much a client may be able to pay in any given period accurately predict the risk of a client defaulting Objective Describe the problem according to the stated goals and solve it.

To solve the problem, you will need to find data. At a minimum the data should include:

client file, including diverse demographic data loan data, including loan tenure, amount, payments, loan status, and a loand identifier account transactions

Deliverables

For every loan analyzed, the submission files should contain the predication as well as the predicted contribution on a monthly basis.

Solution Approach

We will do following steps:

1. Load application_train.csv
2. Check dataset and do Data Cleaning by dropping columns with most missing values since our model will not learn anything from such features/variables
3. Exploratory Data Analysis and check for good possible predictors
4. Calculate **<u>Weight of Evidence (WOE) and Information Value (IV)</u>** of each predictor
5. **<u>Data Imputation</u>**

   We can do data imputation by using various methods such as

   Use WOE values and replace the column values with appropriate woe values. This will handle the null values as well.

   Use FancyImputer to impute missing values

   Use SimpleImputer to impute missing values with median/mode

   Manually replace missing values with median or mode by data visualisation techniques

   Replace null or infinite values with 0

   Here we will replace missing values with value 0, this because after data analysis, using binary or multi-variate analysis we found that data is missing due to some reason like car age is null because the flag own car is 0 i.e. no car, similarly days of employed is 365243 for pensioners or unemployed people. Even though we can replace these values with FancyImputer techniques but that imputation will also not be relevant to the fact. So we choose to replace the null or infinite values with 0 for ease of assignment.
6. Data Preparation by doing Data Transformation for categorical columns using LabelEncoder

Solution Approach

Part - I (with only application_train.csv)

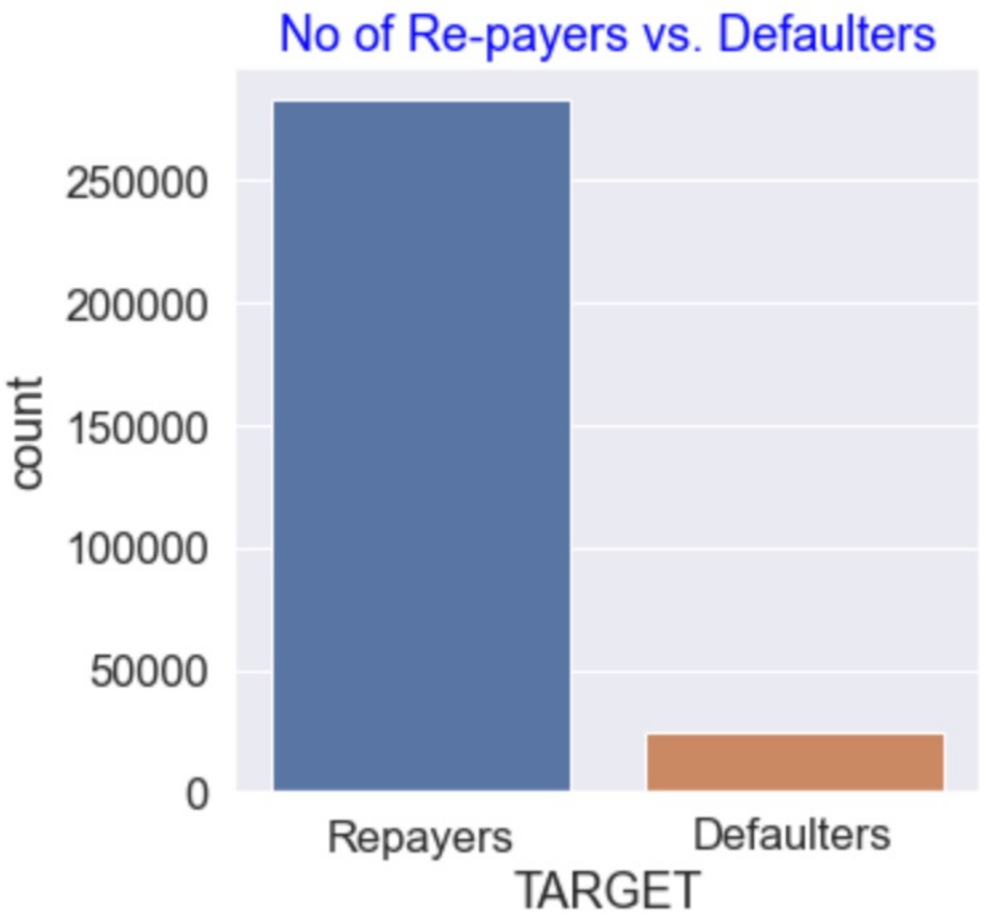
1. Create X and y and do Stratified fold to create X_train and y_train (stratified train-test split)
2. Use RFE and choose first 30 important features /variables
3. Build Stats Model and check for VIF of variables
   Drop variables with high VIF
4. Build Logistic Regression Model with filtered features from above step.
5. Build Random Forest Model with all features
6. Build Random Forest Model using SMOTE balancing algorithm to balance target class,
7. Build XGBoost Model
8. Do Model Evaluation for train-test accuracy, specificity, ROC-AUC
9. Check for probability cut-off value to get stabilised Specificity value
10. List important features
11. Choose the best model (Random Forest in this case)
12. Build the final model
13. Check for feature Importance
14. Build Lift and Gain chart and check for top% of defaulters for 80% default rate

Solution Approach

Part - II (with all files like application_train.csv, previous_applications.csv, bureau data and installments data)

1. aggregate previous_applications.csv, bureau data and instalments data by creating data frame with aggregate values. Use max, mean, sum etc. function for aggregation. Aggregate based on SK_ID_CURR which is parent Id
2. Merge these aggregated dataframes with application dataframe
3. Data Preparation by doing Data Transformation for categorical columns using LabelEncoder
4. Create X and y and do Stratified fold to create X_train and y_train (stratified train-test split)
5. Build Random Forest Model with all features
6. Do Model Evaluation for train-test accuracy, specificity, ROC-AUC
7. Check for probability cut-off value to get stabilised Specificity value
8. List important features
9. Create gain and lift charts and calculate the top percentage of clients that can be identified as defaulters with 85% rate
10. Evaluate model for credit loss saved and revenue loss without model

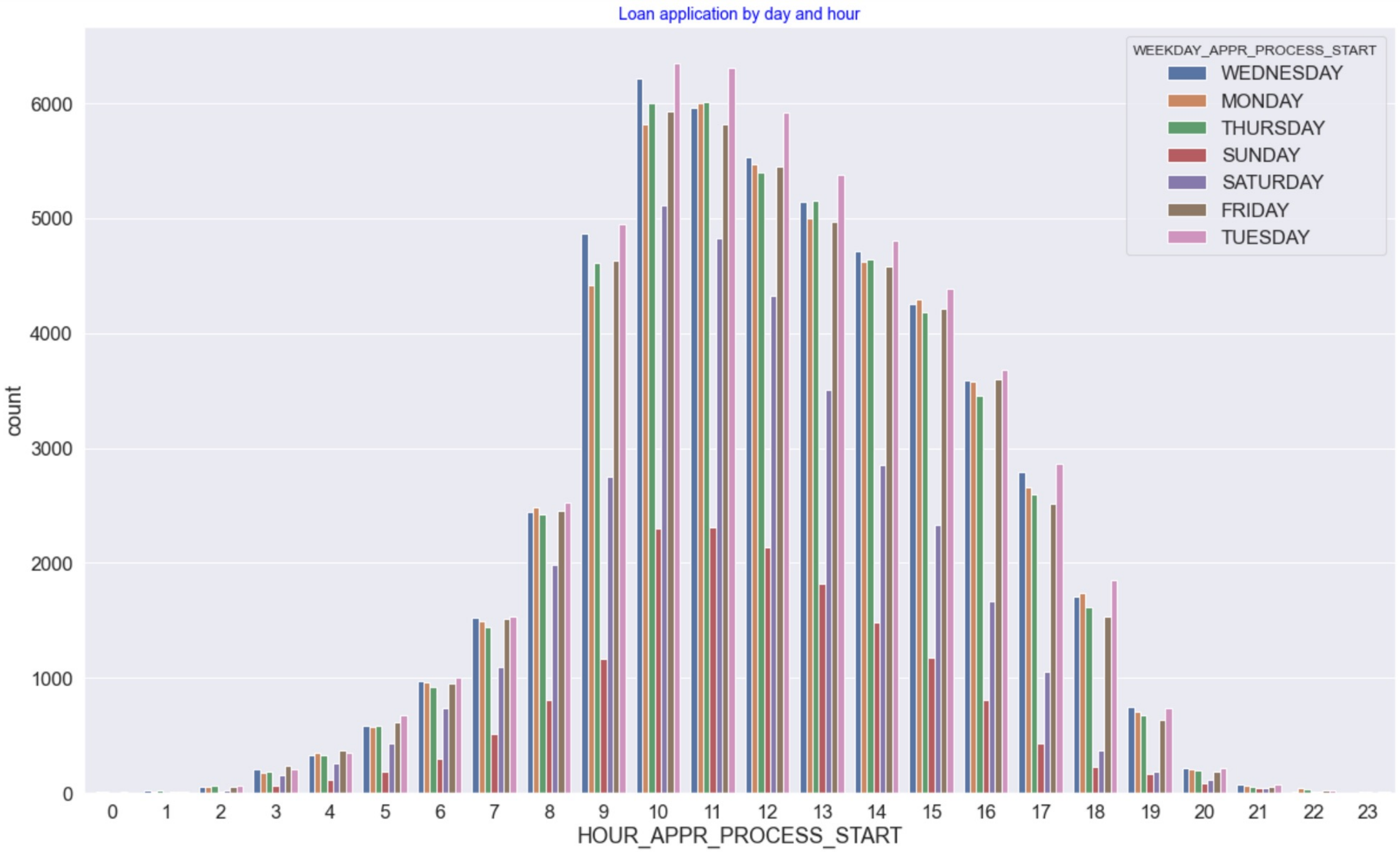
Defaulters and Non-Defaulters in given application data



Observation

- The data is highly imbalanced with only 8.07% of TARGET variable with values as 1
- We will have to handle this during data preparation/modeling

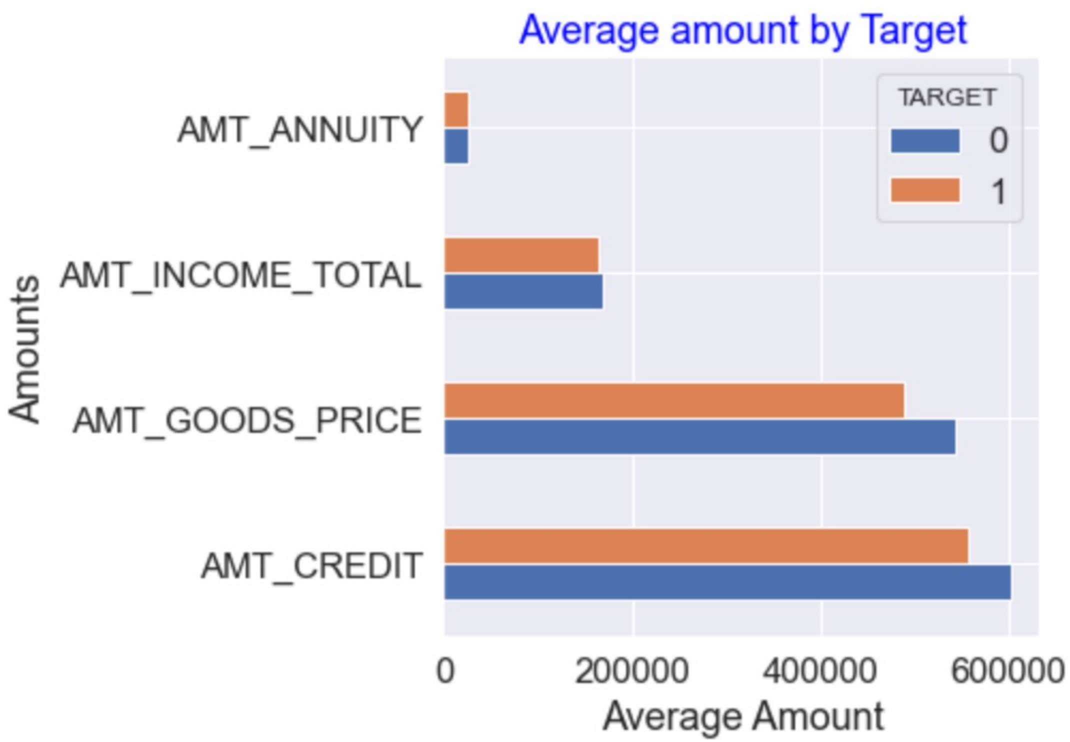
Loan Application by Days and Hours



Observations from above plot

- The loans were applied mainly on Tuesday (17.53%) followed by Wednesday(16.89%)
- 9am - 2pm are peak hours for loan application

Average amounts



Observations from above plot

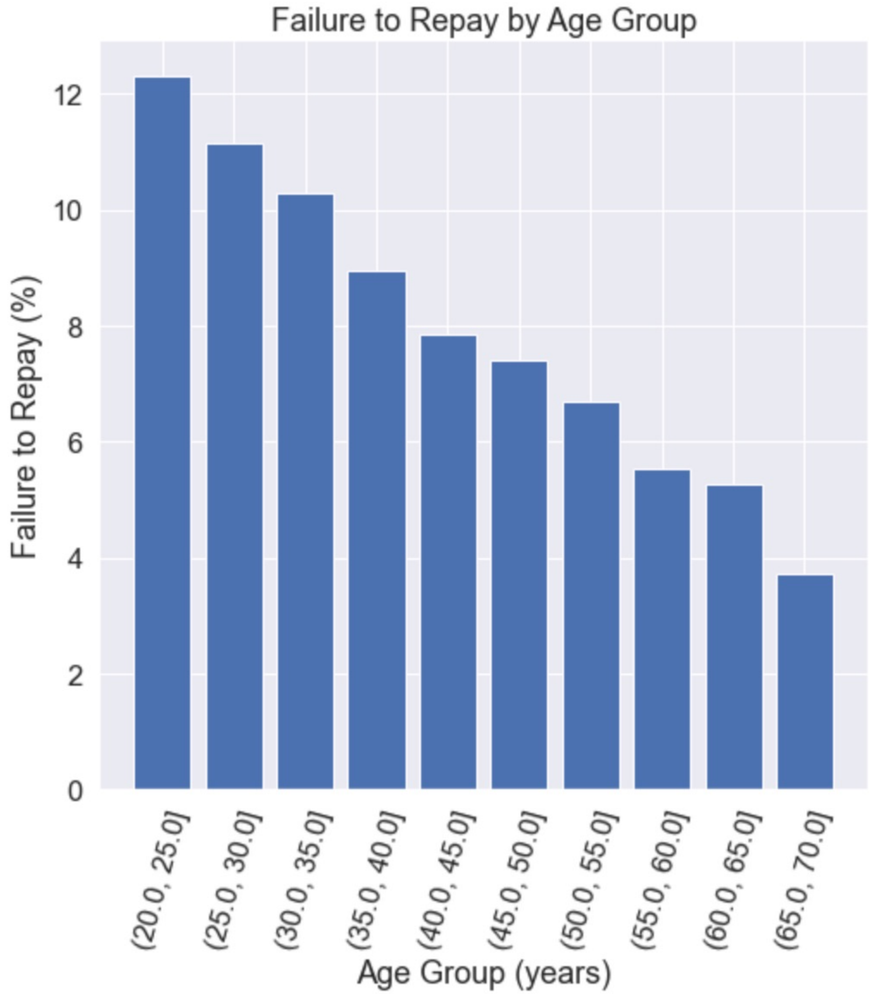
- Target
  - 0 - who pay regularly
  - 1 - who are loan defaulters or have difficulty in loan repayment
- Average annuity amount of people who pay regularly and who default on loan is almost same
- Average income amount of people who pay regularly and who default on loan is almost same
- Average credit amount of people who pay regularly is higher than average credit amount of loan defaulters

Loan Repayment Comparison By Gender

Observations

- There are 3 types of Gender Code - Male, Female and XNA interpreted as Code not available
- Maximum number of loans is taken by Females and maximum defaulters are also females
- In comparison the percentage of repayment difficulty is more by males

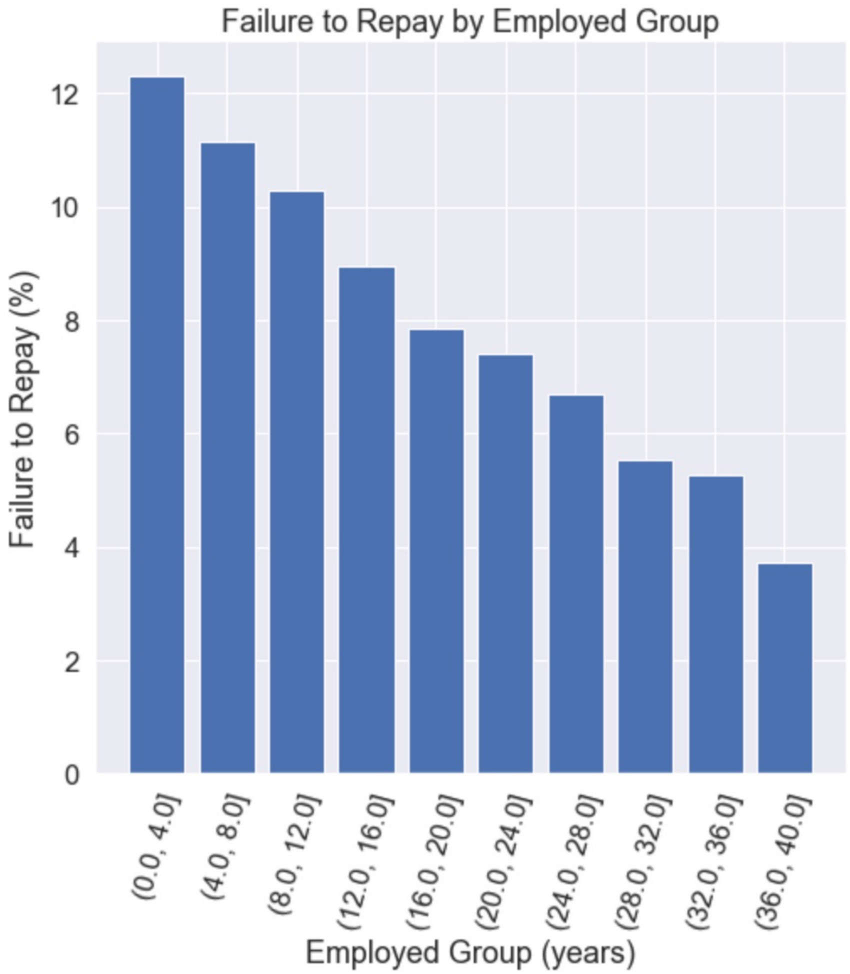
Loan Repayment Comparison By Age Group



Observation ¶

- Majority of young people in the age group of 20-35 are defaulters. The highest defaulter falls in the age group of 20-25
- On the other hand aged people default less. Also count of aged people taking loan is less

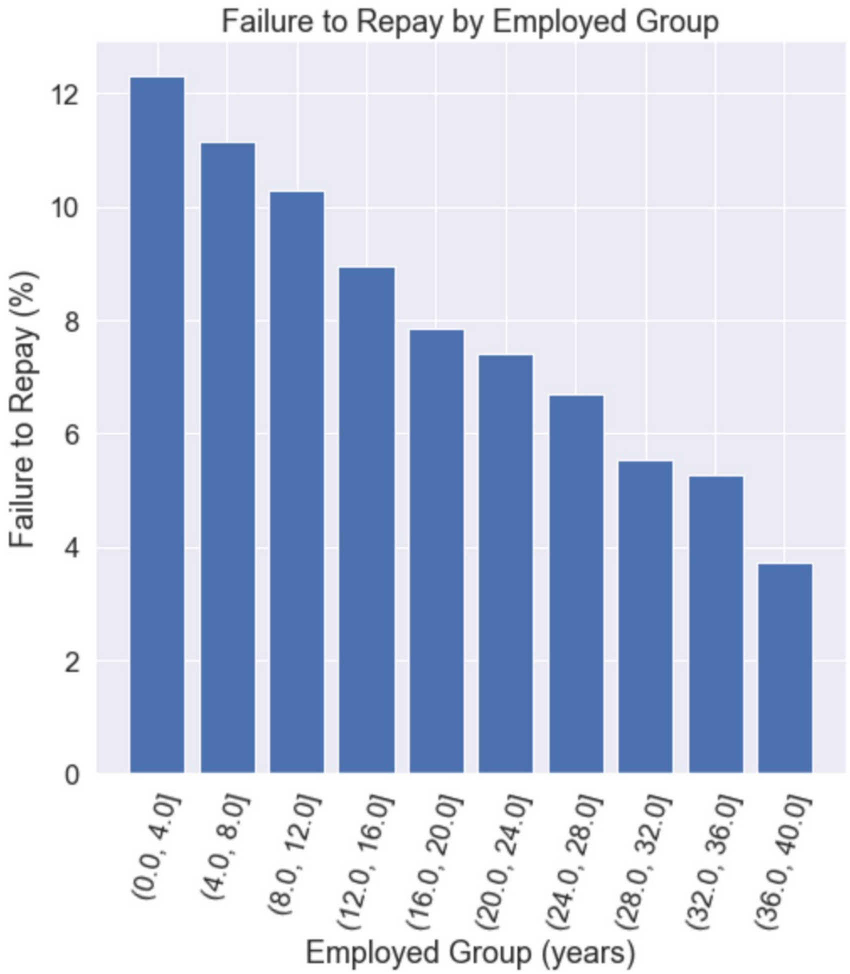
Loan Repayment Comparison By Days Employed



Observation

- Majority of defaulters are in the employed year group of 0-12. The highest defaulter falls in the year group of 0-4. The count of people taking loan in this group is also highest
- On the other hand people who have been employed for long default less. Also this group of people take less loan

Loan Repayment Comparison By Days Employed



Observation

- Majority of defaulters are in the employed year group of 0-12. The highest defaulter falls in the year group of 0-4. The count of people taking loan in this group is also highest
- On the other hand people who have been employed for long default less. Also this group of people take less loan

Information Values of features after merging application data with all other data files

	var_name	iv
17	APPS_EXT_SOURCE_MEAN	0.61416
180	EXT_SOURCE_2	0.30325
181	EXT_SOURCE_3	0.29837
143	BU_CREDIT_DEBT_RATIO_MAX	0.14350
144	BU_CREDIT_DEBT_RATIO_MEAN	0.13603
89	BU_ACT_CREDIT_DEBT_RATIO_MAX	0.12782
140	BU_CREDIT_DEBT_DIFF_MAX	0.12756
43	BU_750_CREDIT_DEBT_RATIO_MAX	0.12505
156	BU_DAYS_CREDIT_MEAN	0.12277
86	BU_ACT_CREDIT_DEBT_DIFF_MAX	0.12158
40	BU_750_CREDIT_DEBT_DIFF_MAX	0.12137
141	BU_CREDIT_DEBT_DIFF_MEAN	0.09960
90	BU_ACT_CREDIT_DEBT_RATIO_MEAN	0.09621
18	APPS_EXT_SOURCE_STD	0.09347
87	BU_ACT_CREDIT_DEBT_DIFF_MEAN	0.09102
179	EXT_SOURCE_1	0.08778
172	DAYS_BIRTH	0.08384
241	INS_D365DPD_DAYS_MAX	0.08155
155	BU_DAYS_CREDIT_MAX	0.08121
157	BU_DAYS_CREDIT_MIN	0.07621
22	APPS_INCOME_EMPLOYED_RATIO	0.07461
280	ORGANIZATION_TYPE	0.07337
153	BU_DAYS_CREDIT_ENDDATE_MEAN	0.07135
159	BU_DAYS_ENDDATE_FACT_MEAN	0.07028

As per definition

```
IV             |          Predictive Power
-----------------------------------------
<0.02          |     Useless for Prediction
0.02 to 0.1    |     Weak Predictor
0.1 to 0.3     |     Medium Predictor
0.3 to 0.5     |     Strong Predictor
>0.5           |     Suspicious or too good to be true
```

Observations from above cell

- Going by Information Value, feature **'APPS_EXT_SOURCE_MEAN'** seems to be too good to be true. We will still consider it to be a good predictor.
- EXT_SOURCE_2 seems to be **strong predictors**
- EXT_SOURCE_3, BU_CREDIT_DEBT_RATIO_MAX, BU_CREDIT_DEBT_RATIO_MEAN, BU_ACT_CREDIT_DEBT_RATIO_MAX, BU_CREDIT_DEBT_DIFF_MAX seem to be **medium predictors**
- APPS_EXT_SOURCE_STD, EXT_SOURCE_1, APPS_EMPLOYED_BIRTH_RATIO, DAYS_BIRTH, APPS_INCOME_EMPLOYED_RATIO, OCCUPATION_TYPE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, DAYS_LAST_PHONE_CHANGE, CODE_GENDER, DAYS_ID_PUBLISH, AMT_GOODS_PRICE, DAYS_REGISTRATION seem to be **weak predictors**

We will build various models and check for model evaluation metrics and feature importance

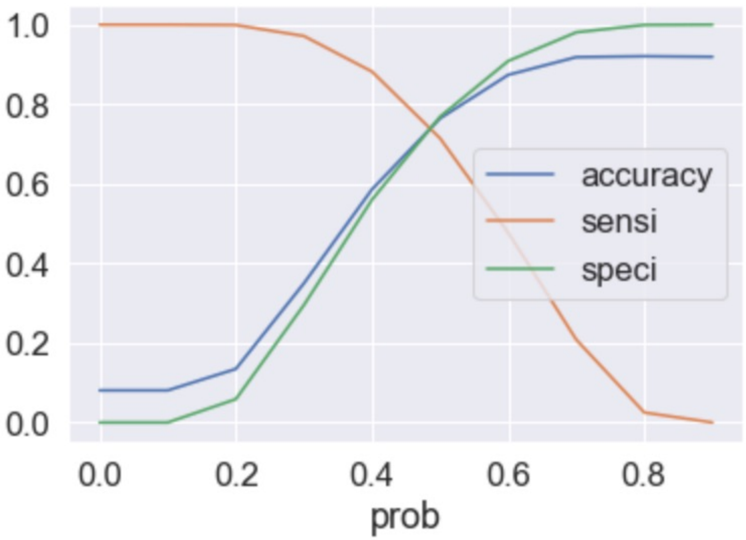
Build various models for Loan Default Prediction

	model_type	train_acc	train_sen	train_spec	train_f1	test_acc	test_sen	test_spec	test_f1
0	StatsModel	0.76	0.58	0.77	0.28	0.75	0.56	0.77	0.27
1	Logistic_RFE	0.75	0.59	0.76	0.27	0.74	0.57	0.76	0.27
2	RandomForest	0.79	0.61	0.8	0.32	0.78	0.54	0.8	0.28
3	RandomForest_SMOTE	0.76	0.46	0.79	0.24	0.76	0.41	0.79	0.21
4	XGBoost_Smote	0.68	0.64	0.69	0.25	0.68	0.6	0.69	0.23
5	RandomForest_AllFiles	0.83	0.6	0.85	0.36	0.81	0.49	0.84	0.3

Model Conclusion

- RandomForest with all files gave better specificity score. We will use this model as our final model to calculate the credit scores

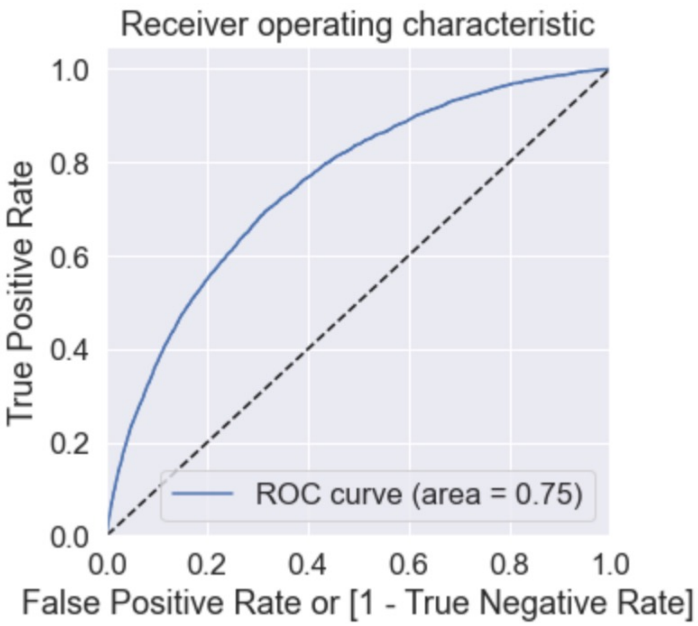
Metrics used for modeling



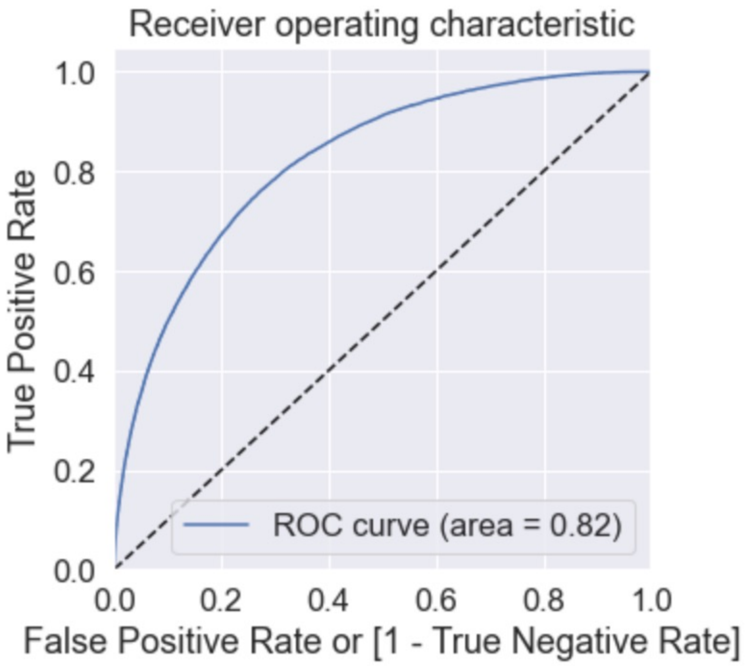
Observations from above plot

- 0.55 is the optimum point to take it as cutoff probability

ROC Curve – Train data

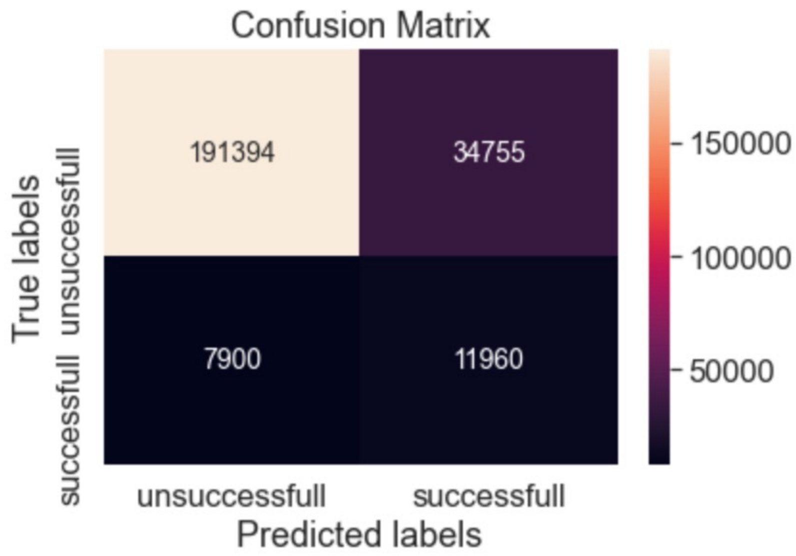


ROC Curve – Test data



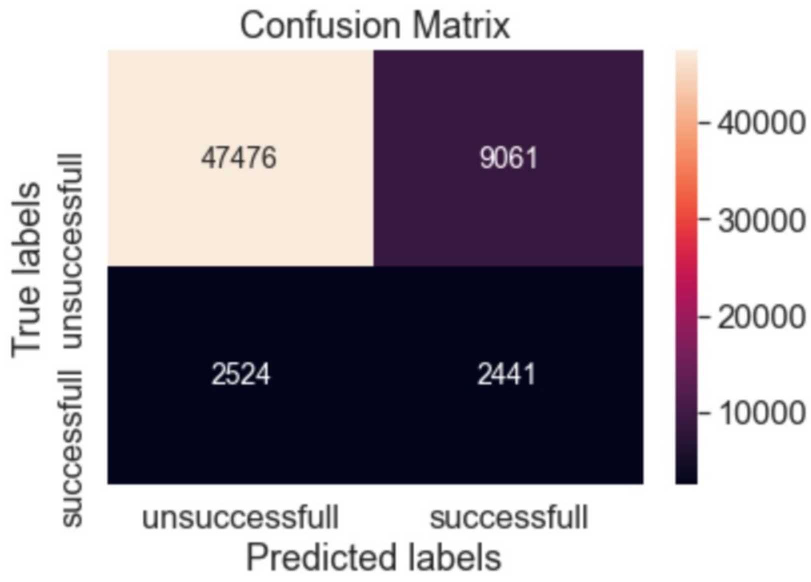
Metrics used for modeling

Confusion Matrix – Train data



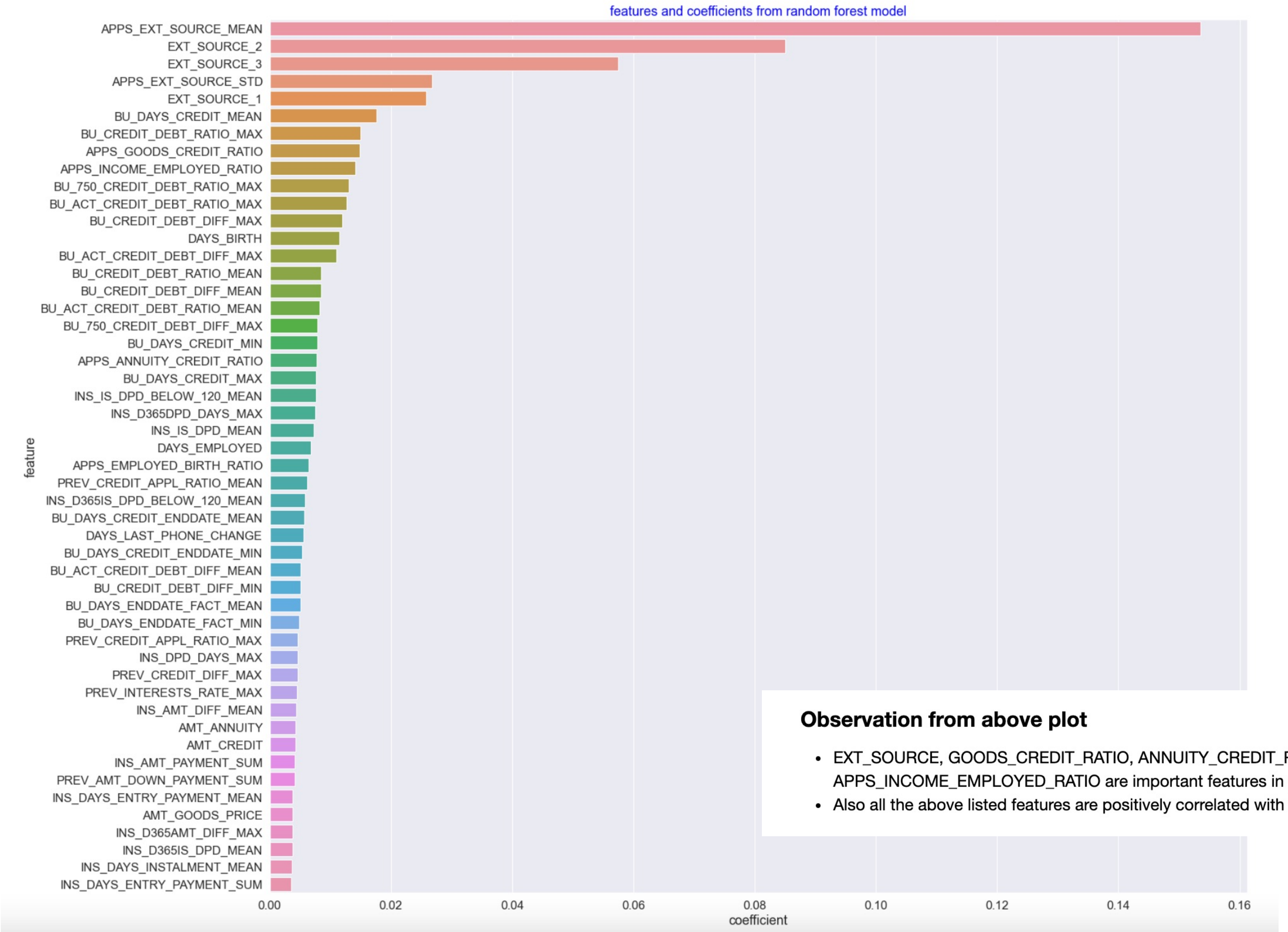

```
Accuracy score: 0.8266120345190623
Sensitivity score: 0.6022155085599195
Specificity score: 0.8463181353886154
f1-score: 0.35929402929027415
Precision score: 0.2560205501444932
Recall score: 0.6022155085599195
AUC: 0.82
```

Confusion Matrix – Test data




```
Accuracy score: 0.8116321420441611
Sensitivity score: 0.49164149043303124
Specificity score: 0.8397332720165556
f1-score: 0.2964717313414708
Precision score: 0.21222396105025212
Recall score: 0.49164149043303124
AUC: 0.75
```

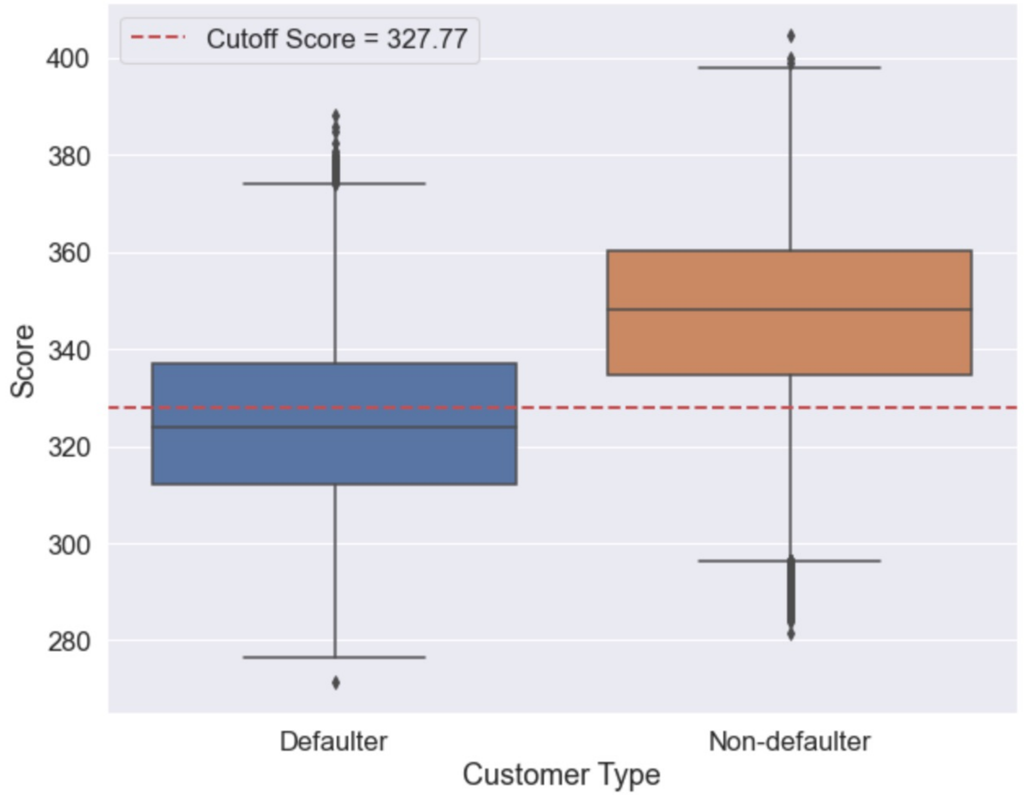
Feature Importance



Observation from above plot

- EXT_SOURCE, GOODS_CREDIT_RATIO, ANNUITY_CREDIT_RATIO, BU_750_CREDIT_DEBT_RATIO_MAX, BU_DAYS_CREDIT_MEAN, APPS_INCOME_EMPLOYED_RATIO are important features in predicting loan default.
- Also all the above listed features are positively correlated with the TARGET

Credit Score of Defaulters and Non-defaulters



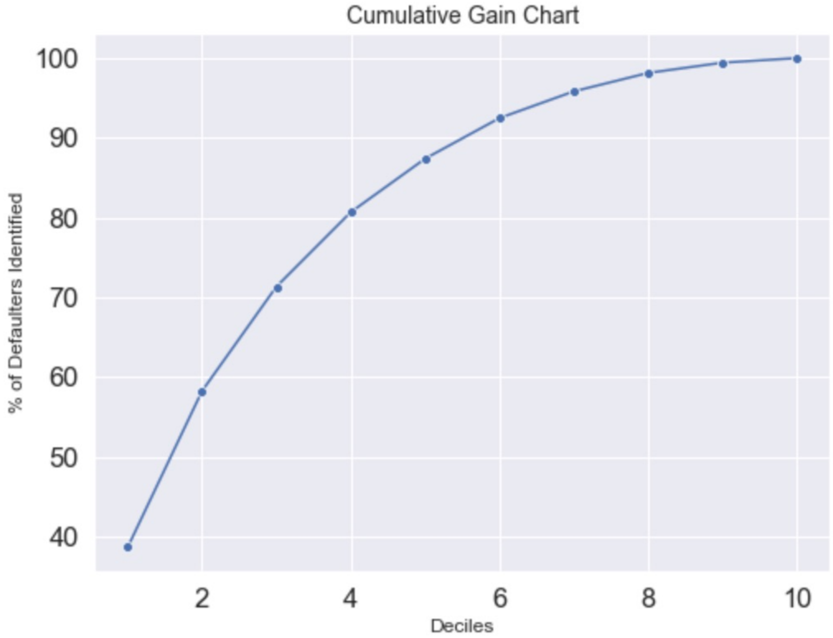
Predicted probabilities and Credit Score

	SK_ID_CURR	TARGET_Prob	TARGET	predicted	Score
0	100002	0.608943	1	1	320.782915
1	100003	0.301166	0	0	357.849266
2	100004	0.248231	0	0	365.533573
3	100006	0.347419	0	0	351.751028
4	100007	0.458261	0	0	338.390033

Observation

- Cut-ff Credit Score is 327.77
- Mean and median of Non-defaulter is higher than the defaulters
- There are some outliers in defaulters with high credit score as well

Cumulative Gain



Conclusion and Recommendation

- EXT_SOURCE, GOODS_CREDIT_RATIO, ANNUITY_CREDIT_RATIO, BU_750_CREDIT_DEBT_RATIO_MAX, BU_DAYS_CREDIT_MEAN, APPS_INCOME_EMPLOYED_RATIO are important features in predicting loan default.
- The cutoff Credit score is 327.77
- Total number of defaulters with Credit Score more than or equal to cut-off score: 10424
- Total number of defaulters with Credit Score less than cut-off score: 14401
- We can conclude that we can predict 87% of the total defaulters by analysing only 50% of the client base

Credit Loss saved

```
Total no. of customers who are actual defaulters and predicted as non-defaulters with model: 10424
Total Actual defaulters : 24825
Total Customers: 307511
% of candidates approved and then defaulted when model was not used : 8.07%
% of candidates approved and then defaulted when model was used : 3.39%
% of Credit Loss saved : 4.68%
```

Revenue Loss saved

```
Total no. of customers who are actual non-defaulters and predicted as defaulters with model: 43816
Total Actual non-defaulters : 282686
Total non-defaulters correctly identified by model : 238870
% of good customers identified by our model : 15.50%
```

Evaluating Financial Benefits of the Model

We will make some assumptions regarding the average credit loss for each defaulted customers and the profit obtained from each non-defaulters.
We will analyse the overall financial benefit of the model and calculate the net financial gain obtained by using the model.
Let's assume the average credit loss for each defaulted customer is CURR 100000/- and profit for each non-defaulters be CURR 10000/-

```
Net profit without model : Curr 34.44 Million
Net profit with model : Curr 134.63 Million
Net Financial gain using the model : Curr 100.19 Million
% Financial Gain : 290.96%
```