PROJECT REPORT

- By Tavishi Singh

1. Problem Statement

The dataset given to us represents period data for US loans taken with unspecified uniform periods. As in the real world, loans may originate before the start of the observation period (this is an issue where loans are transferred between banks and investors as in securitization). The data contains various attributes of mortgage-level-loans which occured in the history as well as the end of the loan for the bank.

Objectives:-

1. Data pre-processing & Exploratory Data Analysis for relations between attributes.
2. Bank Perspective: Predicting whether a loan will be paid-off or defaulted from the bank's perspective.
3. Customer Perspective: Determining whether it's better to continue paying a loan (even in unfavourable conditions) or to default early as which would have happened eventually and save money as given to the bank.

2. Data Explanation

Our dataset Contains 6,00,000 entries over 50,000 loans having default, pay_off and loan continuance.

Feature	Type	Description
id	Nominal	Borrower ID
time	Nominal	Time stamp of observation
orig_time	Constant	Time stamp for origination
first_time	Constant	Time stamp for first observation
mat_time	Constant	Time stamp for maturity
balance_time	Time Series Real Value	Outstanding balance at observation time
LTV_time	Time Series Real Value	Loan-to-value ratio at observation time, in %
interest_rate_time	Time Series Real Value	Interest rate at observation time, in %
hpi_time	Time Series Real Value	House price index at observation time, base year = 100
gdp_time	Time Series Real Value	Gross domestic product (GDP) growth at observation time, in %
uer_time	Time Series Real Value	Unemployment rate at observation time, in %
REtype_CO_orig_time	Binary	Real estate type condominium = 1, otherwise = 0
REtype_PU_orig_time	Binary	Real estate type planned urban development

		= 1, otherwise = 0
REtype_SF_orig_time	Binary	Single-family home = 1, otherwise = 0
investor_orig_time	Constant	Investor borrower = 1, otherwise = 0
balance_orig_time	Constant	Outstanding balance at origination time
FICO_orig_time	Constant	FICO score at origination time, in %
LTV_orig_time	Constant	Loan-to-value ratio at origination time, in %
Interest_Rate_orig_time:	Constant	Interest rate at origination time, in %
hpi_orig_time	Constant	House price index at origination time, base year = 100
default_time	Binary	Default observation at observation time
payoff_time	Binary	Payoff observation at observation time
status_time	Nominal	Default (1), payoff (2), and non default/non payoff (0) observation at observation time

3. Data Preprocessing

3.1 Data Cleaning

270 Null Values were found in the LTV_time feature column so we removed the corresponding entries from the dataset.

3.2 Data Structuring

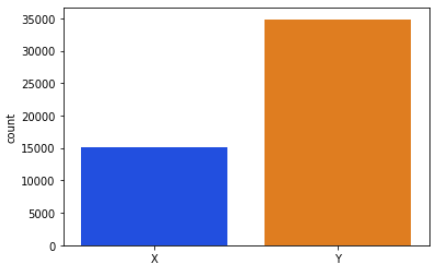
To analyse our **objective 2 and objective 3** we created two different tables. For the **Bank Perspective objective**(i.e. Objective 2) we grouped the data by id feature column and took the last entry from each group thus creating a dataset table with **49982 entries and 17 feature columns**. The status_time feature column is the target column for this objective. All the non default/non payoff cases were considered as payoff because of their data distribution resemblance to payoff data entries. This dataset table is used for binary classification task that determine whether the customer will pay off the loan (status_time = 1) or default the loan (status_time = 0).

For the **Customer Perspective Objective** we took all the entries with the time series real value data type features thus creating a dataset table with **6,00,000 entries and 7 feature columns.**

3.3 Outlier Analysis And Removal

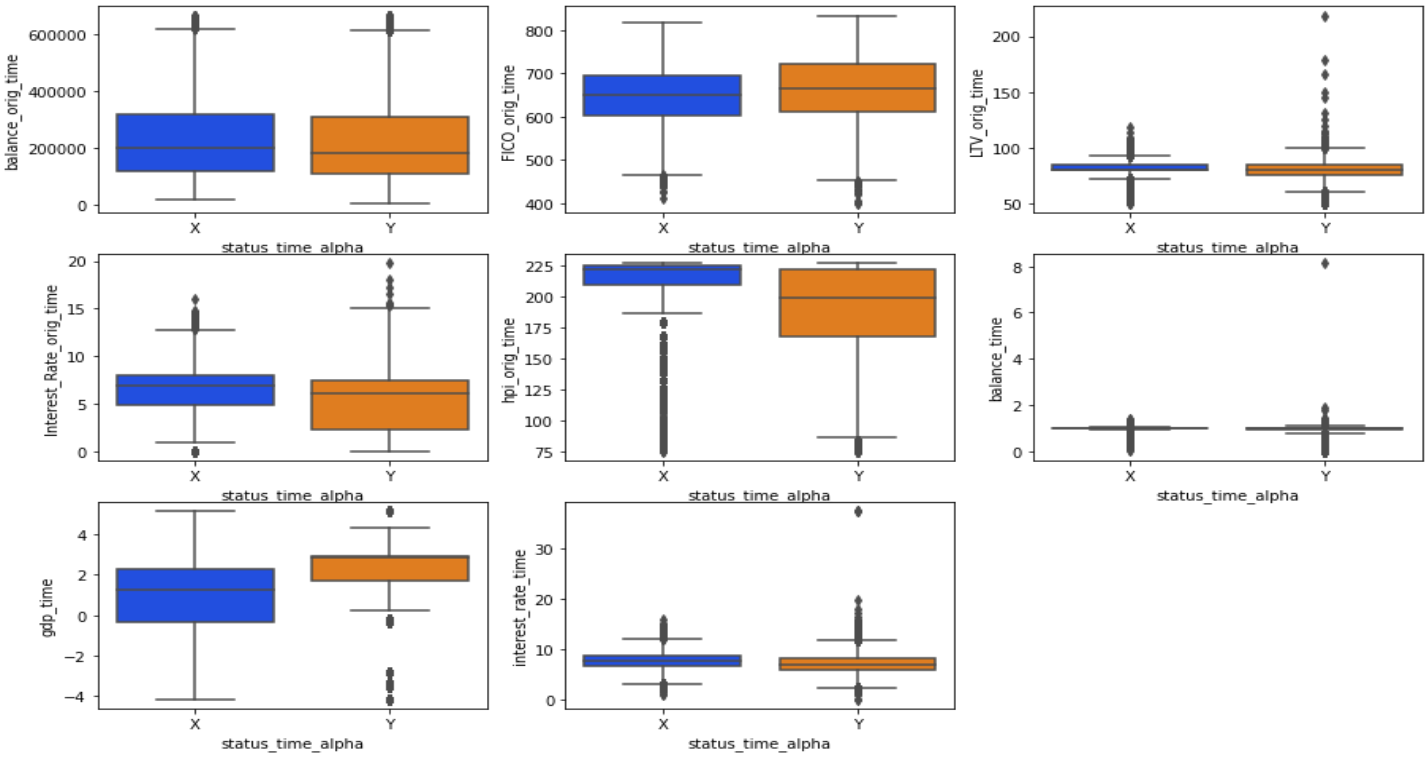
The IQR test for outliers on the balance_orig_time feature of binary classification dataset gave 1728 outliers and so they were removed from the dataset.

3.4 Exploratory Data Analysis



There is an imbalance in the binary classification dataset as out of 49982 entries 15149 are default cases and 34833 are payoff cases. So we applied SMOTE oversampling to remove the imbalance but results

obtained were not satisfactory so we discarded the SMOTE oversampling and continued with our original dataset without oversampling.



Observations from the Above Box Plots for Binary Classification Dataset:

1). Median of gdp_time distribution is higher for payoff cases as compared to that of default cases.

2). Median of LTV_orig_time distribution is lower for payoff cases as compared to that of default cases.

3). Interest_rate_orig_time is lower for payoff cases as compared to that of default cases.

4). hpi_orig_time is lower for payoff cases as compared to the default cases

5). FICO_orig_time is comparable for both the classes with median slightly higher for the payoff cases

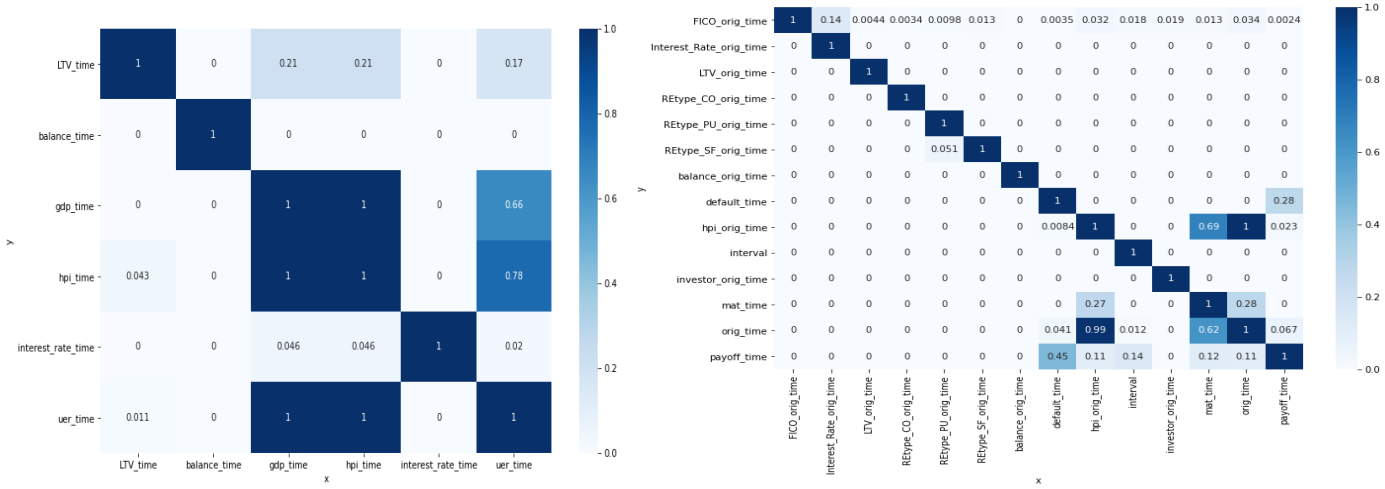
3.5 PCA:

We have 17 features in our binary classification dataset so we applied **PCA** in order to do the **Dimensionality Reduction**.

We set the value of n_components to be 0.95 as we want the **explained variance** to be between **95–99%** and based on this value we obtain in **total 9 principal components (features)**. So we **reduced** our **number of features from 17 (in our original dataset) to 9 factors (PCs)**. We then applied all the classification models on both the PCA features and Original Binary Classification Dataset features and compared the performance of both.

3.6 PPS Matrix:

Performing a single PPS for all the features would give us inaccurate predictions therefore we have performed it on individual datasets.



From the first PPS heatmap we can see that there is a **strong relation** between the “uer_time” and “hpi_time”, “gdp_time” and “hpi_time”, “uer_time” and “gdp_time”. So keeping only one feature out of the 3 will be good enough for further analysis.

From the second PPS heatmap, we can see that there is a strong **bijective** relation between “orig_time” and “hpi_orig_time” and therefore the “orig_time” was dropped.

3.7 Scaling and Normalization

We have used **Standard Scaling** (centering around mean with unit variance) in our static variables (fixed for one ID). In our time series data we have performed **Min-Max Scaling** on “balance_time”.

Some of the features were skewed and we employed **Yeo Johnson** transformation but the results obtained with preliminary analysis were worse.

4. Modelling

Objective: Bank Perspective: This problem was a binary classification problem(Defaulted vs Non-Default). We have used SVC (a non-linear kernel based) , Decision Tree (tree based model without ensemble modelling) , LIghtGBM(tree based **with ensemble modelling)** , ANN(a dense network) models.

For SVC and ANN, **One Hot Encoding** was used and with Decision Tree and LightGBM **Label Encoding** was used. We have hyperparameter tuning using **Grid Search CV** and **k-fold** cross validation with **9-1 train-test split**.

Objective: **Customer Perspective**: We have used the model obtained in the above problem which gives us a **probability to default**. Therefore, The problem was to predict the probability in the future of

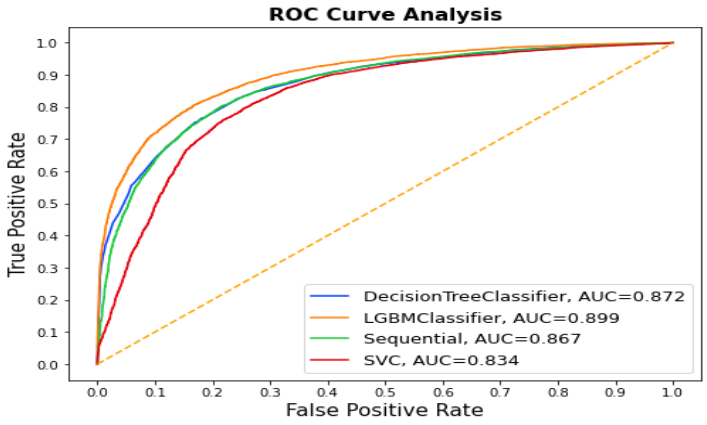
defaulting, for this we need to predict the value for 6 variables (a **multi-variate time-series** prediction). **VAR** and **VARMAX** were chosen as they perform better in multi-variate scenarios. We have chosen **9-1 train-test split** for the validation purposes.

For each customer the next prediction of probability of being was obtained and the decision can be taken whether to default or not based on next period's prediction.

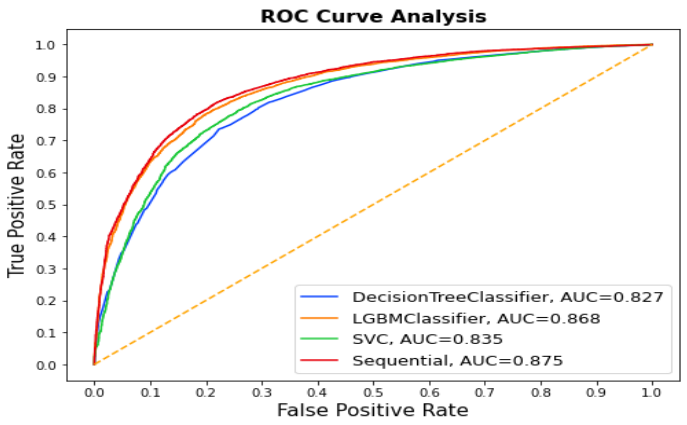
5. Results

(Objective: Bank's Perspective) **For original features (Graph - 1)**

Model	Details	Accuracy	Precision	Recall	ROC AUC score
Decision Tree	criterion = entropy max_depth = 7	0.81	0.84	0.9	0.872
LightGBM	max_depth = 3 n_estimators = 100 Learning_rate = 0.3	0.84	0.87	0.9	0.889
SVC	C = 1 kernel = linear	0.8	0.83	0.91	0.834
Neural Network	3 hidden layers with 64, 32, 16 hidden nodes respectively, 2 dropout layers (0.1), Trained for 100 epochs with batch_size = 64 activation = relu	0.81	0.86	0.87	0.867



Graph -1 (ROC for original features)



Graph-2 (ROC for PCA features)

(Objective: Bank's Perspective) **For PCA features (Graph - 2):**

Model	Details	Accuracy	Precision	Recall	ROC AUC score

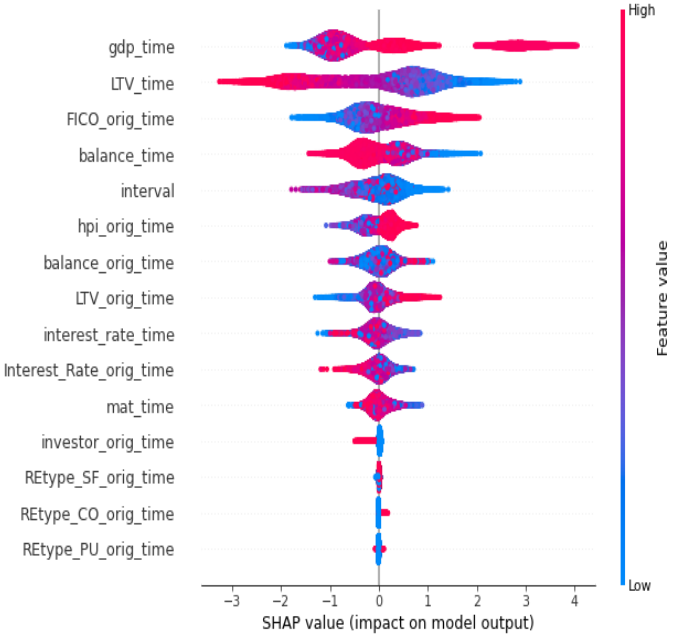
Decision Tree	criterion = gini max_depth = 7	0.79	0.83	0.88	0.827
LightGBM	max_depth = 3 n_estimators = 100 Learning_rate = 0.3	0.86	0.88	0.89	0.87
SVC	C = 1 kernel = linear	0.78	0.79	0.93	0.835
Neural Network	3 hidden layers with 64, 32, 16 hidden nodes respectively, 2 dropout layers (0.1), Trained for 100 epochs with batch_size = 64 activation = relu	0.82	0.87	0.87	0.875

For the **original features** we obtain the highest AUC value 0.899 for **LGBM classifier(highest)** followed by that of decision tree, Neural Network, SVC (Linear) respectively in decreasing order of AUC values.

For the **PCA features** we obtain the highest AUC value 0.875 for the Neural Network classifier followed by that of LGBM, SVC (Linear), decision tree respectively in decreasing order of AUC values.

Highest accuracy of **0.86** as well as highest precision of **0.88** was obtained for the LGBM classifier using PCA features and the highest accuracy of **0.84** was obtained for LGBM classifier using original features.

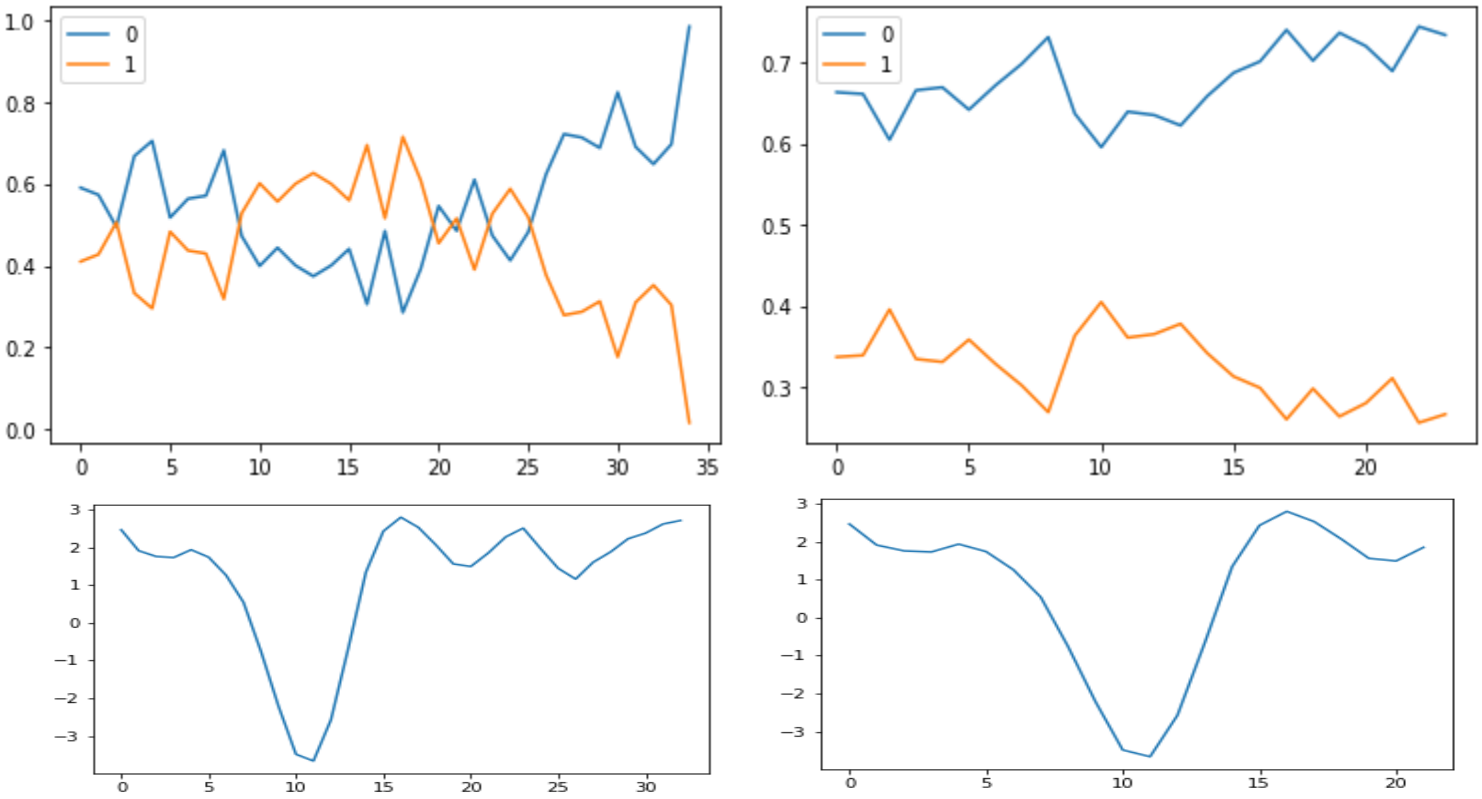
In terms of accuracy LGBM with PCA features performed better as compared to every other model and in terms of **ROC_AUC score LGBM with original features performed the best among all the models.**



This graph summarises the impact (positive or negative) on the model output based on the feature value of features. Take gdp_time as eg. we can observe that the feature value (i.e. Red color darkening) is higher on the positive x axis which means that higher gpd_time value will have a positive impact on the model output and lead to the customer paying off his/her loan.

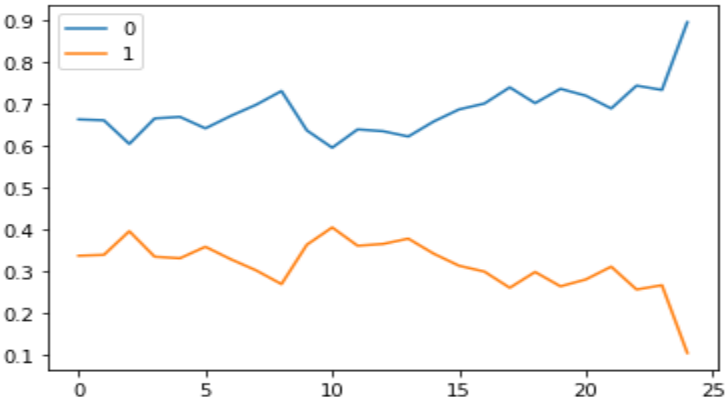
For LTV_time we can observe that the feature value (i.e. Red color darkening) is higher on the negative x axis which means that higher LTV_time value will have a negative impact on the model output and lead to the customer defaulting his/her loan.

(Objective : Customer's Perspective):



The top-left graph shows the **probability of non-defaulting** in a customer's case **(ID:A)**who **paid off** his loan. Here(bottom left) we can see that the weighted average mean(rolling mean) at the last period is showing an increase in the future.

The top-right graph shows the **probability of non-defaulting** in a customer's case**(ID:B)** who **did not** pay off the loan. Here(bottom left) we can see that the weighted average mean(rolling mean) at the last period is showing an increase in the future which predicts that the person should have **not defaulted**(by comparing with the above case).



In case of **customer B,** on predicting the output feature from the time series models from the time series models and predicting the probability from the model obtained in first part we get that the probability of defaulting decreased a lot (orange line) which means that the person **should not have defaulted and would have gotten the property instead of dropping the loan.**

Percentage of Customers who have defaulted but could have been prevented according to market conditions by this model are 0.3979 % out of 15149 loans.