Capstone Project

Credit Card Default Prediction

Content

Problem Statement

- This project is aimed at identifying the customers who would default on their credit card payments next month in Taiwan.

- Default is failure to make the credit card payments on time.

- The objective is to build a supervised classification model which gives the best predictive accuracy of the probability of the defaulters, which has good degree of separation between the two classes.

Overview of the data

Attributes

- LIMIT_BAL - Amount of the given credit (NT dollar)
- Gender (1 = male; 2 = female)
- Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- Marital status (1 = married; 2 = single; 3 = others)
- Age (year)
- Repayment status in last 6 months
- Amount of bill statements in last 6 months
- Amount of payment in last 6 months
- default payment next month (Yes = 1, No = 0)

Dependent Features

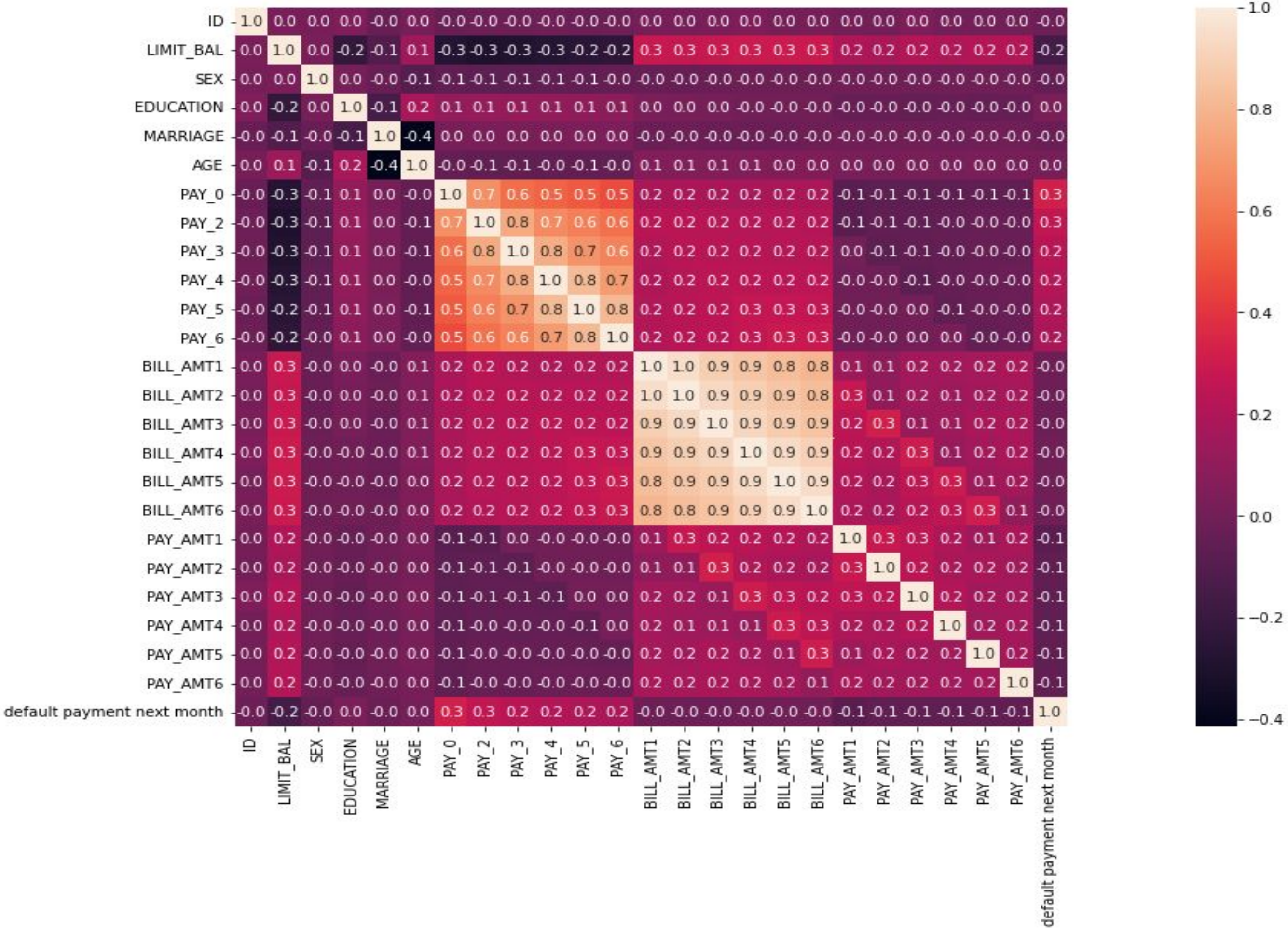
Independent Feature

EDA

Actions on the dataset

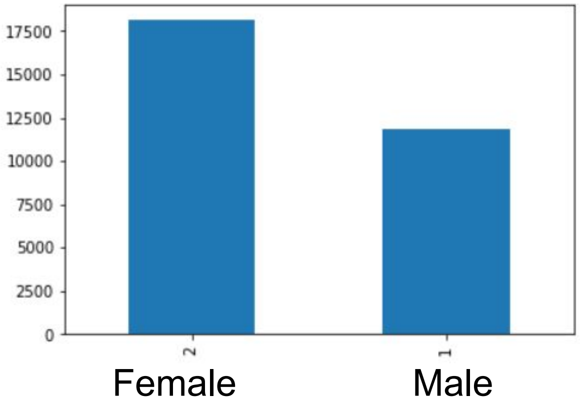
- The dataset is of 30000 entries and 25 columns
- There are no null values
- There are no categorical variables in the data. Variables likes Education, Sex, Marriage were already encoded to numerical variables.
- Values of different columns are of different range , so scaling is required – MinMaxScaler is used to normalise the data.

Feature Dependency

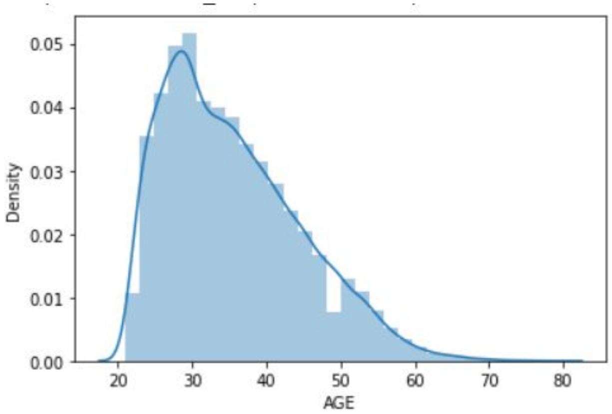


Observations from the plots

Sex

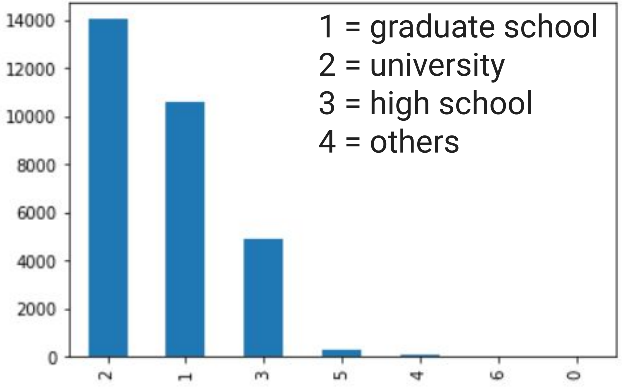


Age

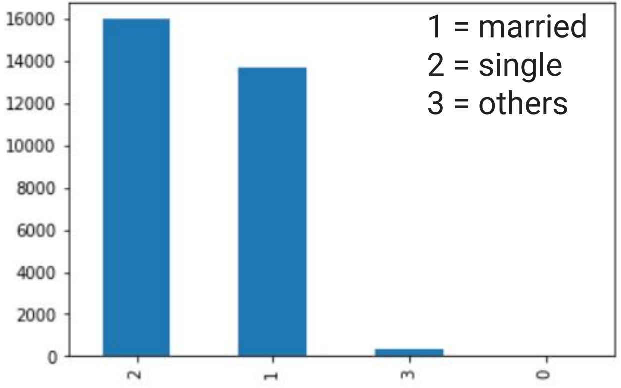


- Female credit holders are more than male credit holders.
- Most of the clients are of the age between 25 to 40.

Education

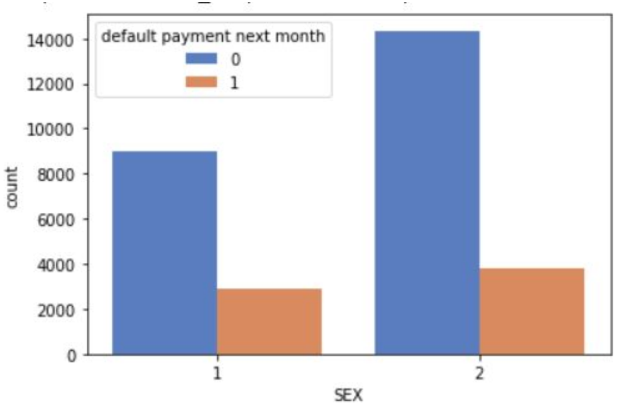


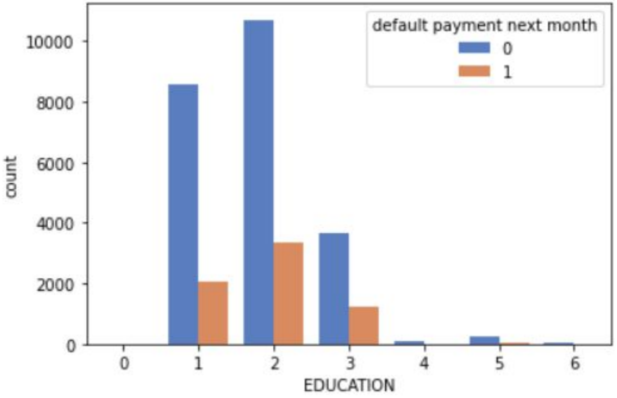
Marital status



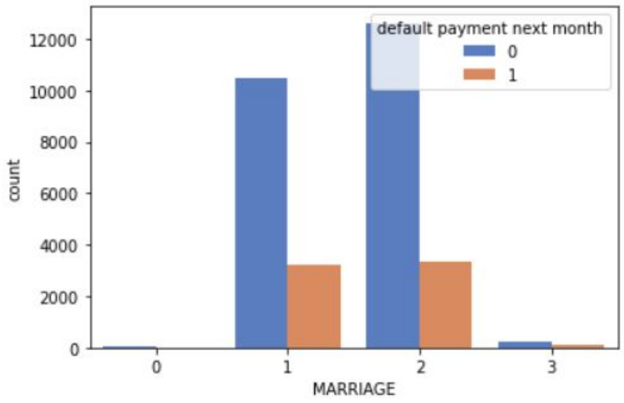
- Most of the clients completed university level studies.
- Number of clients who are single are more than clients who are married.

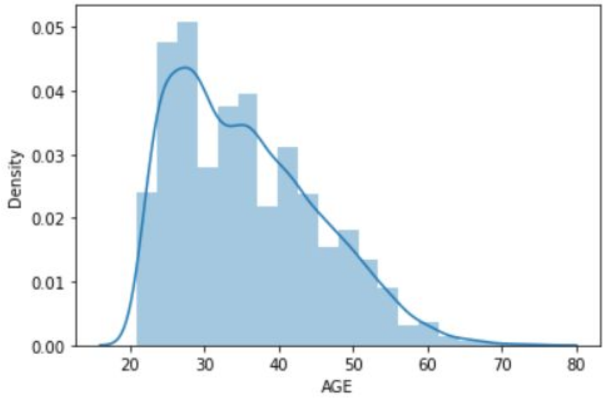
Feature distribution with respect to output variable





- Percentage of male defaulters is higher than females.
- Most of the clients who default their next month pay are from university.



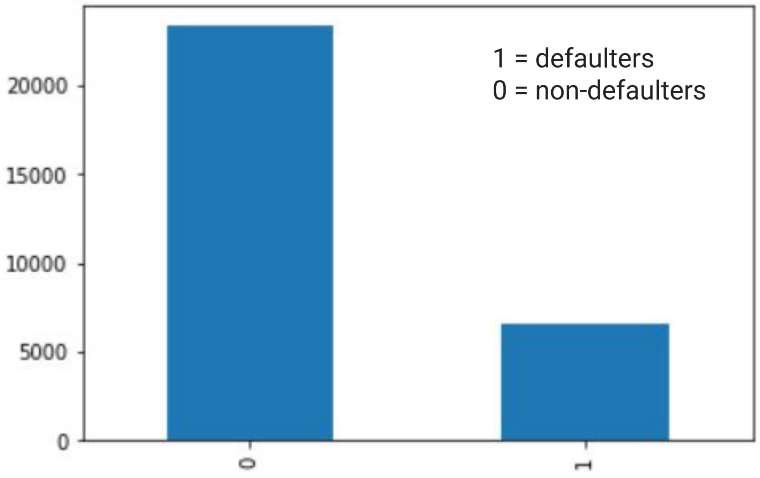


- Percentage of defaulters is more in married people than singles.
- Most of the clients who default their payments are of age between 25 to 35.

Imbalanced data

- Observations in class 0 is much higher than observation in class 1
- **SMOTE** – oversampling process by generating virtual training records to handle the imbalanced class distribution.
- **Synthetic Minority Oversampling Technique** increases the percentage of minority class.
- Total number of observations:
  Original dataset - 30000
  Resampled dataset - 46728

Numbers of defaulters and non-defaulters coming month



Applying models

	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	XGB Classifier	0.875883	0.934345	0.808450	0.866850	0.875849
1	SVC	0.774806	0.864006	0.652013	0.743187	0.774744
2	DecisionTreeClassifier	0.812540	0.806154	0.822723	0.814354	0.812545
3	RandomForestClassifier	0.883016	0.916874	0.842278	0.877994	0.882996
4	KNNclassifier	0.792496	0.805883	0.770340	0.787711	0.792485

Model Evaluation : KS static

- Kolmogorov-Smirnov static measures performance of classification models.
- It is a measure of the degree of separation between the positive and negative distributions(defaulters and non-defaults respectively in our case).
- The K-S is 100 if the scores partition the population into two separate groups in which one group contains all the positives and the other all negatives.
- If model cannot differentiate between positives and negatives, that is if model selects cases randomly from population, the KS will be 0.

Cont..

- The KS value ranges from 0 to 100, higher the value, better the model performance.

- The K-S test measures the distance between the plotted distribution functions of two classifications.

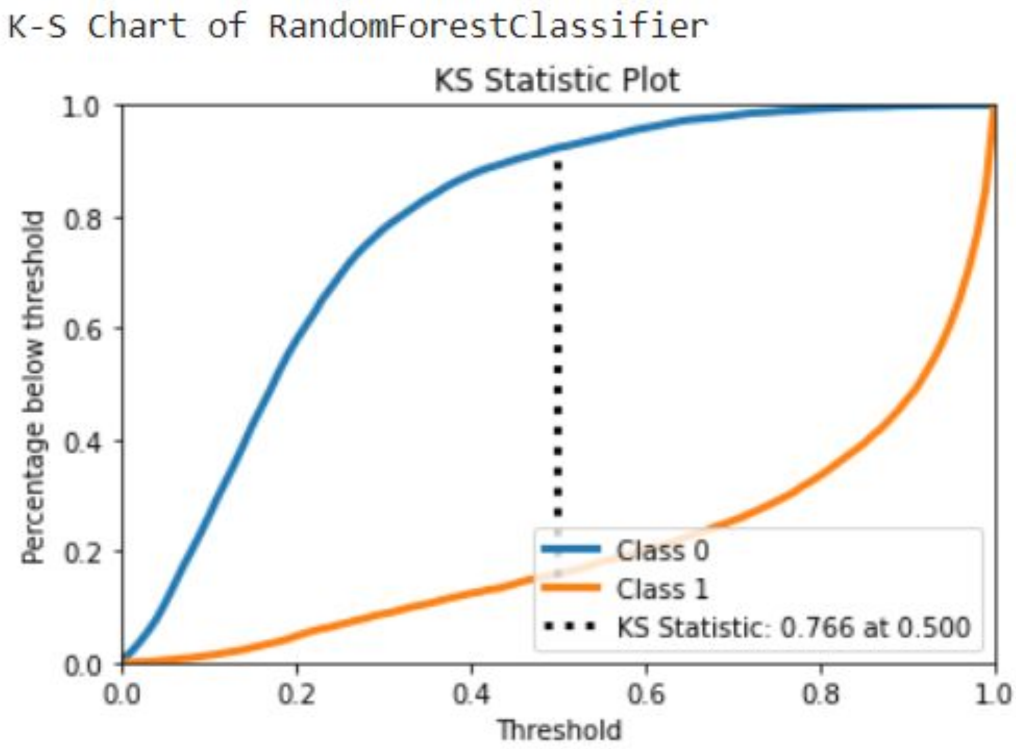
- Each classification score can be transformed to lie between 0 and 1.

- The score that generates the greatest vertical separability between functions is the KS score and the model with high KS score is the best model.

KS table of Random Forest Classifier

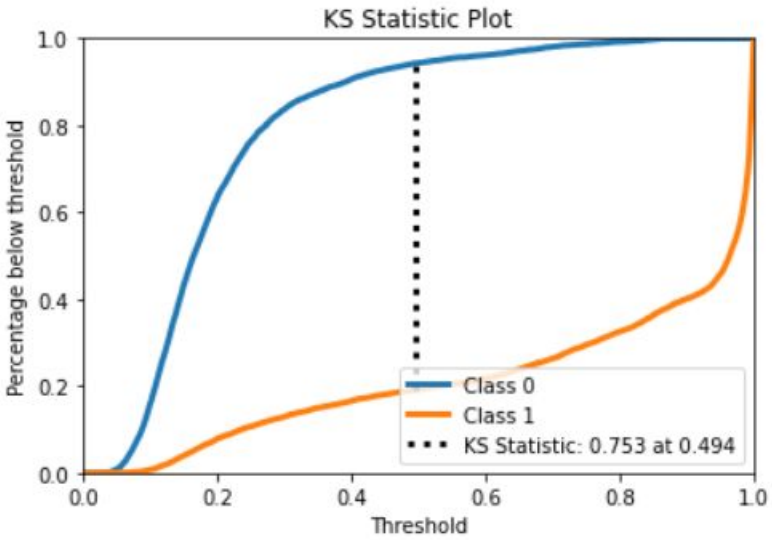
Decile	min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	cum_eventrate	cum_noneventrate	KS
1	1.00	1.00	1058	0	15.10%	0.00%	15.10%	0.00%	15.1
2	0.96	0.99	1677	1	23.94%	0.01%	39.04%	0.01%	39.0
3	0.87	0.95	1422	18	20.30%	0.26%	59.33%	0.27%	59.1
4	0.66	0.86	1261	168	18.00%	2.40%	77.33%	2.67%	74.7
5	0.42	0.65	704	646	10.05%	9.21%	87.38%	11.88%	75.5
6	0.28	0.41	353	1039	5.04%	14.82%	92.42%	26.69%	65.7
7	0.21	0.27	200	1092	2.85%	15.57%	95.28%	42.26%	53.0
8	0.15	0.20	171	1308	2.44%	18.65%	97.72%	60.92%	36.8
9	0.09	0.14	101	1343	1.44%	19.15%	99.16%	80.07%	19.1
10	0.00	0.08	59	1398	0.84%	19.93%	100.00%	100.00%	0.0

KS Chart of Random Forest Classifier

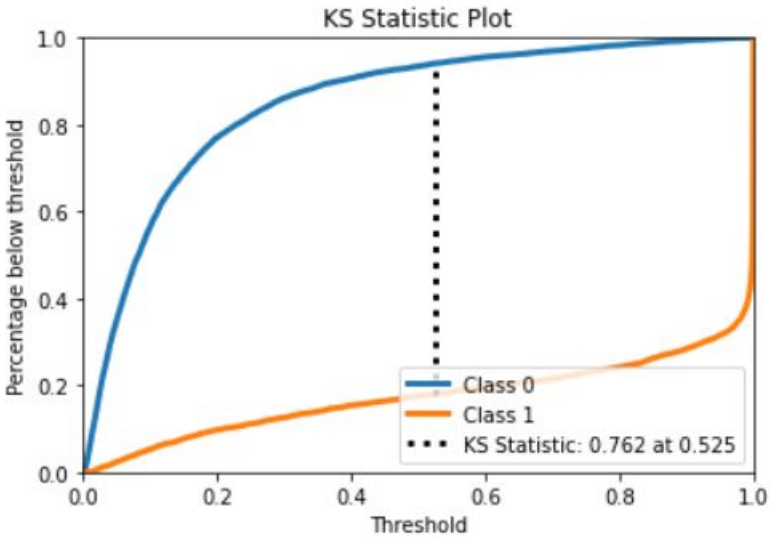


KS Chart of XG Boost Classifier

Before tuning parameters



After tuning parameters



KS table of XG Boost Classifier with best parameters

KS is 75.0% at decile 4

Decile	min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	cum_eventrate	cum_noneventrate	KS
1	0.999832	0.999993	1401	0	20.00%	0.00%	20.00%	0.00%	20.0
2	0.999259	0.999832	1403	0	20.03%	0.00%	40.02%	0.00%	40.0
3	0.992604	0.999259	1402	0	20.01%	0.00%	60.03%	0.00%	60.0
4	0.736257	0.992582	1225	177	17.49%	2.52%	77.52%	2.52%	75.0
5	0.313153	0.736015	661	740	9.43%	10.55%	86.95%	13.08%	73.9
6	0.165582	0.312717	330	1072	4.71%	15.29%	91.66%	28.36%	63.3
7	0.096058	0.165535	229	1173	3.27%	16.73%	94.93%	45.09%	49.8
8	0.055347	0.096015	172	1230	2.46%	17.54%	97.39%	62.63%	34.8
9	0.026156	0.055346	116	1286	1.66%	18.34%	99.04%	80.96%	18.1
10	0.001356	0.026135	67	1335	0.96%	19.04%	100.00%	100.00%	0.0

Conclusion

- Both Random Forest Classifier and XG Boost are performing well.
- Random Forest Classifier has best performance with respect to all metrics we have used such as ROC, Precision, Recall and KS static score.
- XG Boost Classifier has shown accuracy of 88% with KS of 75 after hyperparameter tuning.
- Random Forest Classifier has shown accuracy of 88% with KS of 75.5 in predicting the defaulters coming month.

Thank you