American Express Campus Analyze This 2020

Final Submission

Team Details

Team Name : XEMA

Name	Campus	Roll No.	Mobile No.	Email Id
Akash Kumar	IITK	170069	7355092930	akashkr@iitk.ac.in
Aditya Singh	IITK	170056	7678638709	adisingh@iitk.ac.in

Estimation Technique Used

Please provide the estimation/modeling technique(s)/approach used to arrive at the solution/equation

Raw Data
- Import train and test data from csv file

Pre-processing and Cleaning
- Converting String formatting to numeric
- Handling NaN data
- Feature selection using greedy elimination method based on features-target correlation and pairwise cross correlation
- Plotting correlation matrix amongst features

Stratified Data Split
- Data Split into training and validation set in the ratio 80:20

Undersampling & Oversampling
- Handling problem of class imbalance in the dataset.
- Combining oversampling and undersampling using SMOTE with 0.3 and Random Under Sampler with 0.6 sampling strategies

Data Training Models
- **Gradient Boosting classifier**
- **Random Forest classifier**
- **Artificial Neural Network based classifier**

Model Prediction & Selection
- Model is trained on the training set data, and evaluated over the validation set.
- Confusion Matrix is generated and compared
- ROC curves are generated and compared
- Precision, Recall and F1 scores are reported
- Final Model is selected based on comparision on ROC-AUC score

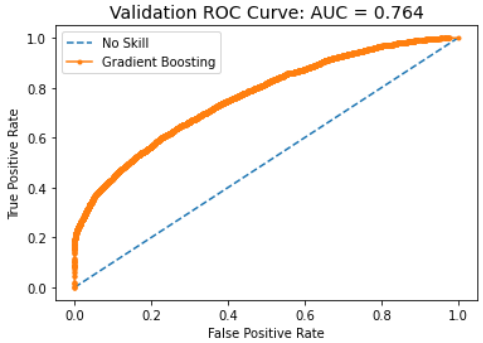
AMERICAN EXPRESS

Strategy to decide final list

Please provide the strategy employed to decide the final list for submission

From the presence of class imbalance, we can infer that accuracy is not the appropriate metrics for evaluation.
ROC AUC score was used to decide the final set of submission, as it represents degree or measure of separability of classes.
High area under TPR and FPR provides a model capability to distinguish the positive label occurrence over negative labels.

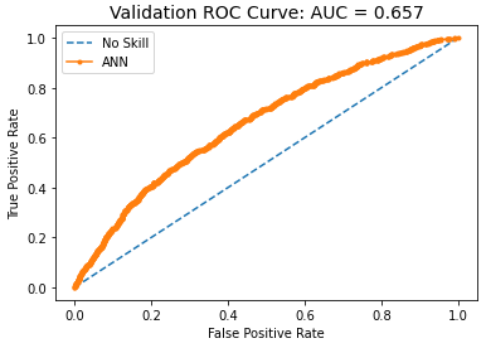
Gradient Boosting



	Precision	Recall	F1-score	Support
0.0	0.72	0.93	0.81	7675
1.0	0.78	0.39	0.52	4785
Accuracy	-	-	0.73	12760
Macro Avg	0.75	0.66	0.67	12760
Weighted Avg	0.74	0.73	0.70	12760

>

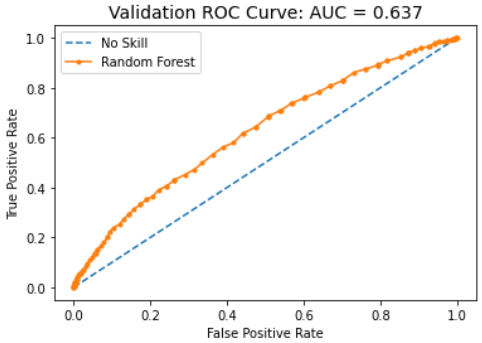
ANN



	Precision	Recall	F1-score	Support
0.0	0.83	0.94	0.88	3190
1.0	0.36	0.14	0.20	729
Accuracy	-	-	0.79	3919
Macro Avg	0.59	0.54	0.54	3919
Weighted Avg	0.74	0.79	0.75	3919

>

Random Forest



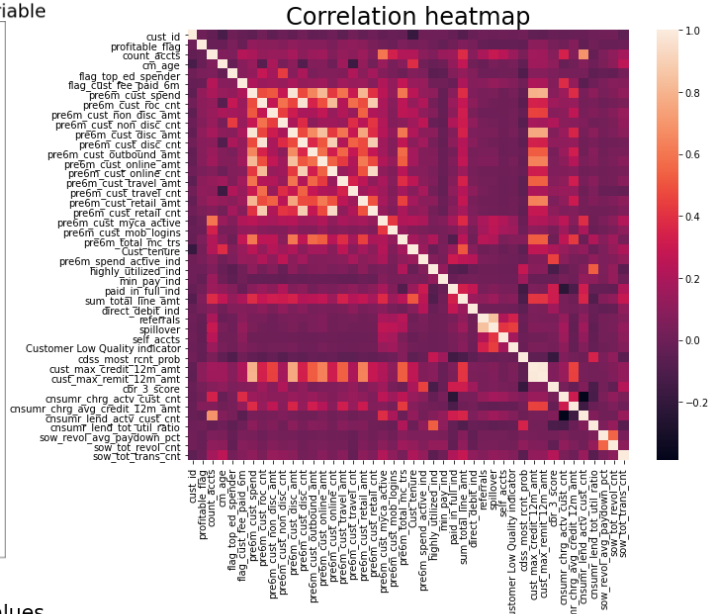
	Precision	Recall	F1-score	Support
0.0	0.83	0.94	0.88	3190
1.0	0.36	0.15	0.21	729
Accuracy	-	-	0.79	3919
Macro Avg	0.59	0.54	0.55	3919
Weighted Avg	0.74	0.79	0.76	3919

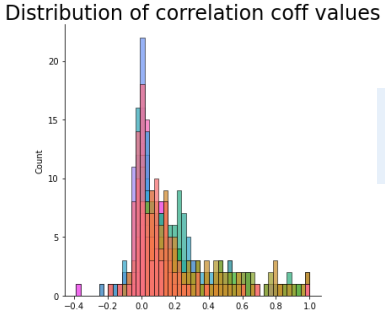
Details of each Variable used in the logic/model/strategy

Please provide details of each variable used in the final logic

Model features used

Counts of acc. number	Binary flag for customer's low quality acquisition
Binary flag for top spending customer	Risk scores for the customer
Binary flag for card fee payment	External Bureau score for customer
Total profit made on customer	Consumer charge avg credit
Total tenure of the customer	No. of active lending card of customer
Binary flag if customer spend on card	Total amount paid over revolving external card balance
Binary flag if customer paid min. card amount	Total spend in last 6 months
Total credit limit of the customer's card	Total transaction counts in last 6 months
Binary flag for auto payment	Max transaction payment made in 6 months
No. of times customer self referred	No. of spill over





Number of features: 49 → 20

AMERICAN EXPRESS

Reasons for Technique(s) Used

Why do you think this is the best technique(s) for this particular problem?

Advantages of Greedy Elimination

Dimensionality reduction and feature selection, dropping features with low target-feature correlation but high pairwise cross correlation. The final features have less multicollinearity among data and are more independent with each other.

Advantages of sampling

Solving problem of class imbalance, which could cause model performance failure. Combination of up sampling and down sampling provides way to regularize the model and prevent overfitting as compared to using only oversampling or under sampling.

Advantages of Gradient Boosted Classifier

Tree based models are more robust to outliers, GBC can also handle null values and ensemble of weak learners build over one another provides more accurate aggregated predictions as compared to linear models which requires feature scaling and are more prone to outliers.

Advantages of ROC-AUC over accuracy

ROC AUC score is equivalent to calculating the rank correlation between predictions and targets. From an interpretation standpoint, it is more useful because it tells us how good at ranking predictions the model is & probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

Final Submission File

Please embed your final submission file (.csv) here.

	cust_id	profitable_flag
0	569986	1
1	569987	0
2	569988	0
3	569989	1
4	569990	0
5	569991	0
6	569992	0
7	569993	0
8	569994	0
9	569995	0
10	569996	0
11	569997	0
12	569998	0
13	569999	0
14	570000	0
15	570001	0

THANK YOU