APPROACH PROPOSAL FOR CREDIT CARD LEAD PREDICTION

For **“Analytics Vidhya - JOB-A-THON”** or **“Happy Customer Bank”**

The approach for the problem is personal and can be similar to existing research. It combines various key concepts information. Feel free to reach me in case of any doubt or clarification

OVERVIEW

This is the problem statement. You can skip if you have the background. Taken from Analytics Vidhya site.

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings. The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc. In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

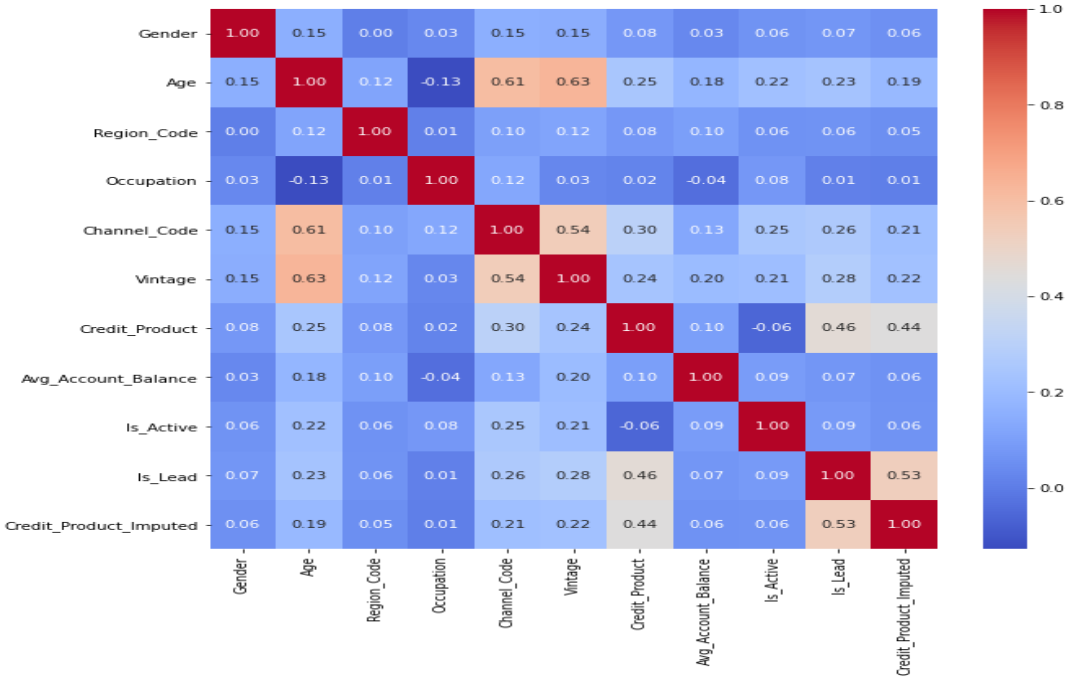
- Customer details (gender, age, region etc.)
- Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.)

The Objective

- Identifying the key person to whom we can pitch. I.e, Customers who can be interested. (Is_Lead)
- Identifying the features that make the cross-sell successful
- Performing Feature Engineering and finding the right solution to the goal

The Process I took

- Step 1: Data Loading and basic data information. I.e, data_type, missing data, EDA
- Step 2: Normalizing the amount column to the log
- Step 3: Conversion of Categorical to numerical using Label Encoder
- Step 4: Filling missing value of column Is Product
    - To replace NAN values with mode value
    - To impute most occurred category and add importance variable **(current)**
    - To create a New Category for NAN Values with “Unknown” label
    - To replace with "Yes" or "No"



The above correlation matrix shows the new variable we created has good correlation with Is_Lead.

- Step 5: Detecting Outliners function (Not used)
- Step 6: Splitting of data into train and valid from train csv file with 33% as valid data and rest as train.
- Step 7: Use of StandardScaler (Not that effective on trial)
- Step 7: Use of Sampling Strategy
    - Under sampling
    - Over sampling
    - SMOTE **(current)**
- Step 8: Experiment with different Machine Learning algorithms

The model that are used and their scores are shown in detail later in the document and training was on train and AUC score on valid. Note Valid data is not the end test data.

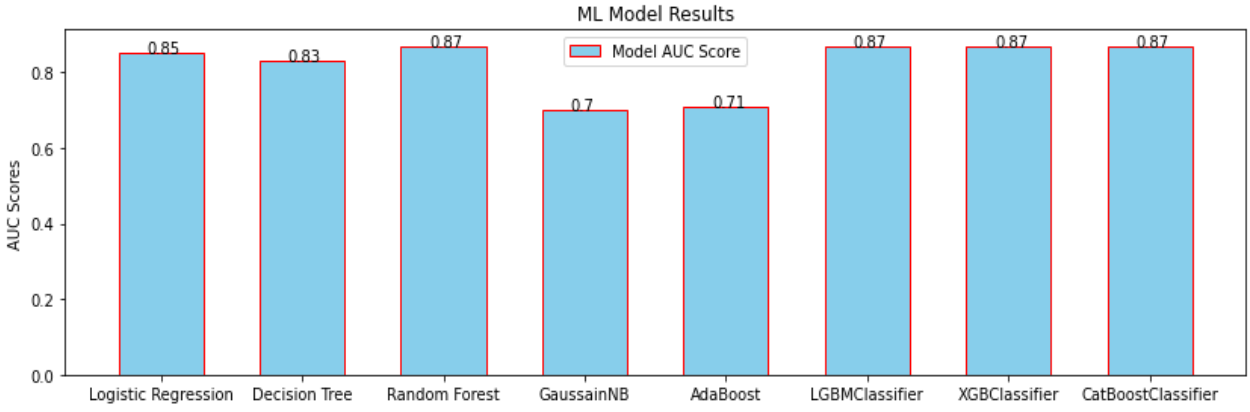
- Step 9: I have also experiment with Tensorflow DNN model. - AUC ~82
- Step 10: Selecting top 3 models with metric of AUC score and hyper tuning
- Step 11: Ensembling of top models
- Step 12: Training on full train + valid data or complete Train file.
- Step 13: For test data I have applied amount normalization and passed for model prediction
- Step 14: Creating Submission file and submitting on site.

Workflow

Experiment 1:

Strategy 1: No sampling strategy and no Standard Scalar.

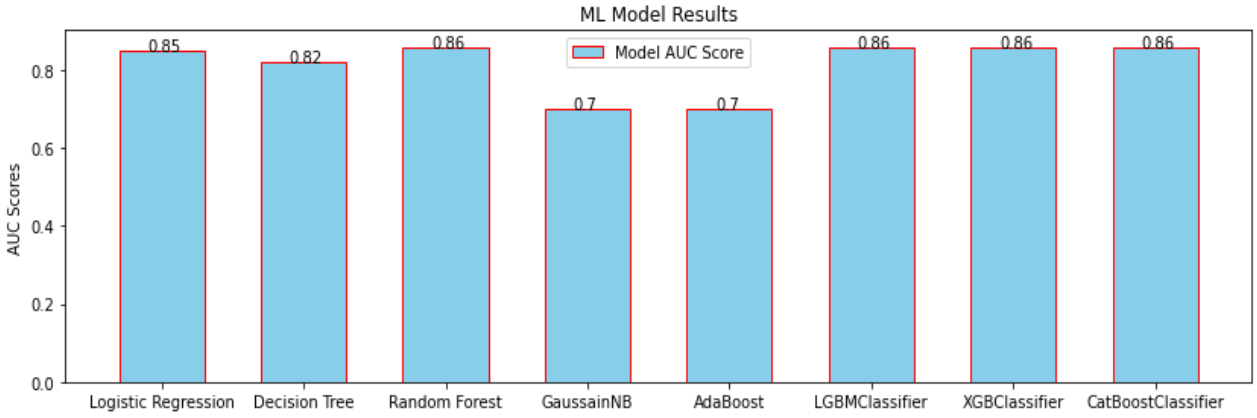
```
Number transactions train dataset:  164635

Number transactions valid dataset:  81090

Total number of transactions:  245725


Train: value 0 means Customer is not interested. Total count 125529

Train: value 1 means Customer is interested. Total count 39106
```



Result on actual Test data on platform: AUC 0.8715

Experiment 2:

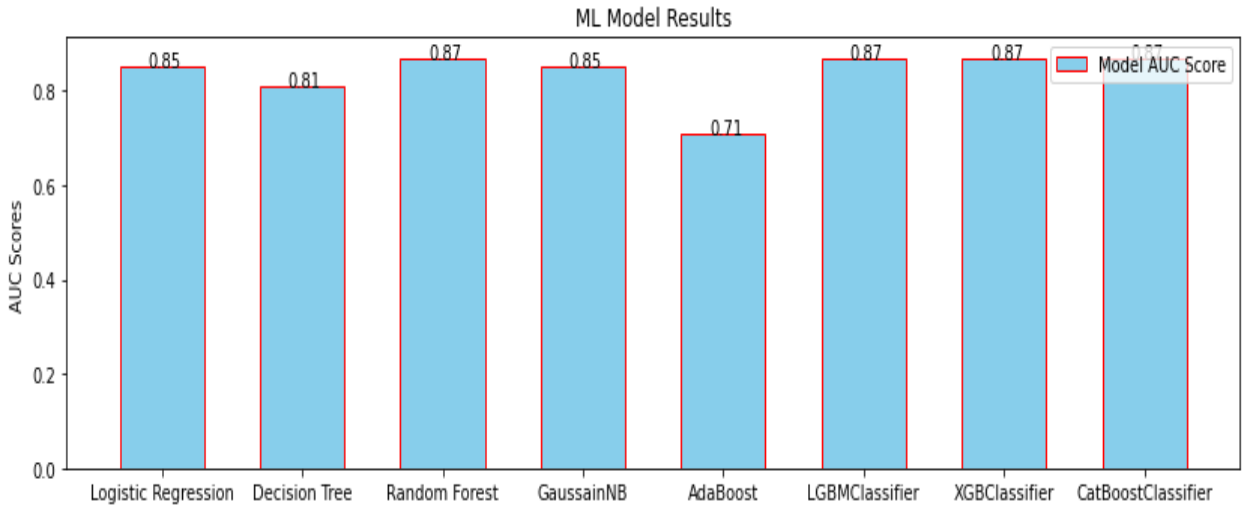
Strategy 2: With Sampling SMOTE and no Standard Scaler



Result on actual Test data on platform: AUC 0.864337690604231

Experiment 3:

Strategy 3: With Standard Scalar and without SMOTE



Strategy 4: Drop Outliers

Strategy 5: Imputing missing data with ML

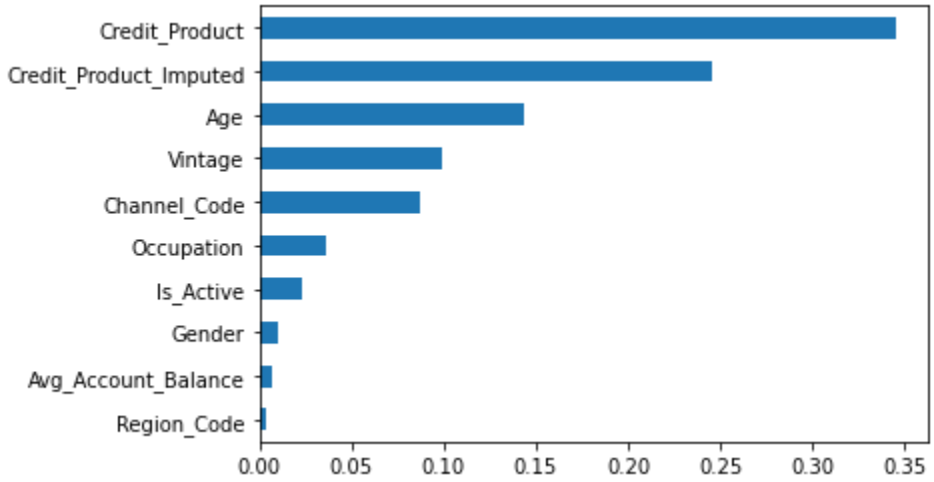
Final Results:

The below results are on X_Valid.

Model Name	Experiment 1 (Strategy 1)	Experiment 2 (Strategy 2)	Experiment 3 (Strategy 3)	Experiment 4* (Strategy 1)
Logistic Regression	0.85	0.85	0.85	0.85
Decision Tree	0.82	0.82	0.81	0.91
Random Forest	**0.87**	**0.86**	0.87	0.87
AdaBoostClassifier	0.70	0.70	0.85	0.7
GaussainNB	0.82	0.7	0.71	1.0
LGBMClassifier	0.88	**0.86**	0.87	0.88
CatBoostClassifier	0.87	**0.86**	0.87	0.89
XGBClassifier	0.87	**0.86**	0.87	0.89
AUC on 30% Test Data	**0.8715**	**0.86433**	**0.8716**	**0.8723**

**refers to model trained on complete data model is prediction*

Features Importance



Future Scope

> *Due to time delay, there were some ideas that I didn't got time for experimenting...*

- Use of Logistic Regression to fill missing data of column "Credit_Product" with and without using Is_lead column
- Use of advanced TensorFlow model for same problem.
- Use of autoencoder to solve the same problem

Learnings

Benefits

- Have learned the importance of class balancing
- Normalization of column
- Ensembling can be very powerful

Mistakes

- For the first few submissions I applied threshold and predicted 0 or 1 and did my submission, while the ask was for the probability score not the label.
- Underestimated the ML model power. I thought using deep neural networks will always be better than traditional ML models. But I was wrong.

CONCLUSION

I look forward to hear from you and supporting the JOB-A-THON to provide the opportunity to prove us and learn from it. I am confident that I can meet the challenges ahead and stand ready to partner with you in delivering an effective solution.

If you have questions on this proposal, feel free to contact me at your convenience by email or by linkedin.

Thank you for your precious time in reading this and please let me know your feedback positive or negative.

Vijender Singh

Email: vijendersingh412@gmail.com

LinkedIn: vijendersingh412
Data Scientist & NLP Expert