AMEXPERT DECIPHER – WOMEN'S MACHINE LEARNING HACKATHON

JENI JAIN

Preliminary Analysis

- From Training Data, we can see that for a given customer, various financial records are given for the three months : **April, May, June.**
- Apart from Age, Gender, Account Type, Region Code, all other 39 columns contains data related to the User Credit and Debit Card Details along with Investments and Loans expenditure.
- Approach to do prediction Credit Card Consumption was to first clean the data , then add the required new features and then build a model on it.
- After initial observations, column “id” was dropped
- The count of Null Values were very significant for a lot of columns. Next step involved data pre-processing to address these issues

Data Pre-processing

1

- There were more nearly 90% Null Values in certain columns. Here, we have assumed that null values means that since the tags are categorical in nature, missing value should be interpreted as **0**.
- **Example:** Null Values in **personal_loan_active** implies that the user has not taken any personal loan.

personal_loan_active	91%
personal_loan_closed	91%
vehicle_loan_active	97%
vehicle_loan_closed	95%

2

- In some variables, like **dc_count_apr**, it was seen that no value was **0,** but the column had high number of Null values. It was also seen that whenever **dc_count_apr** was null, dc_cons_apr was also null – indicating that null cases might highlight no or 0 transactions.
- Correspondingly, null values were replaced with 0 in this cases

dc_count_apr	58%
dc_count_may	52%
dc_count_jun	47%

3

- No such logics were seen to work with certain columns. Columns with very high null values were dropped : investment_1, investment_2, investment_3, investment_4
- Remaining variables less than **17%** data missing and were thus **imputed with mean values**
- Also, some variables were converted to “factor” class for correct analysis in R

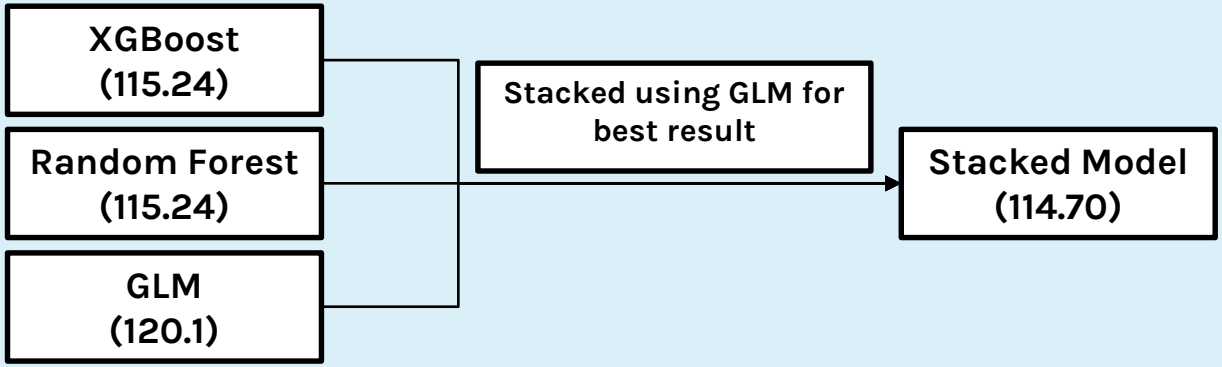
Investment_1	95%
Investment_2	93%
Investment_3	96%
Investment_4	99%

Feature Engineering

- New variables were incorporated based on understanding of target variable.
- Following list includes major variables introduced –
    a. Total Credit Card and Debit Card cons, counts over 3 Months (April, May, June)
    b. Total Credit and Debit amount, counts over 3 Months
    c. Credit Card and Debit Card cons per transaction count for all 3 Months
    d. Credit and Debit amount per transaction count for all 3 Months
    e. Ratio of Credit Card and Debit Card cons for all 3 months
    f. Ratio of Credit and Debit amount for all 3 months
    g. Ratio of Credit and Debit Card cons with every 2 months at a time
    h. Ratio of Credit and Debit amount with every 2 months at a time
- Total of **83 variables** were used for modelling starting with **44 initial variables**

Modelling

- **Data Split**: 75% Training & 25% Test Data.
- **Modification of Target variable:** To directly optimize the RMSE values for the models, the target function was converted to **log(cc_cons + 1).** This allowed to use RMSE as optimization metric while in reality optimizing RMSLE
- **Hyperparameter Tuning:** It was used to find best parameter for each of the following model : XGBoost, Random Forest and GLM.
- **Model stacking** was used to further improve performance over individual models
- Stacking was carried out using GLM and XGBoost, with better results achieved using GLM



Key Takeaways

- Significant number of Null Values in Customer Bank Account Details indicated it to be zero. So, Data imputation for those null values with mean/median was not required.
- **Model Stacking :** Using any of the single ML Algorithm were not able to provide significant RMSLE. Stacking the predictions of various ML Models help us improve RMSLE results.

Points to be Focused for Participants

- Before starting to jump on and start coding, the most important accept is to understand the business case.
- For feature engineering part, the target variable should be kept in mind and try to establish the relationship between it and the new variable.
- Make sure to remove the irrelevant columns from the data.

THANK YOU!