Hello Team 19,

Summary of discussion dated April 10, 2022

Focus:

* Addressing handling missing data
* Identify promising parameters for features Engineering

Participants: Supra, Tessy, Yalini

The meeting was focused on identifying the variables included in the dataset that can be suitable for engineering additional features for model development. Decisions were based on domain knowledge, validating assumptions and deducing the data in specific parameters.

The starting point was finding methods to combine the parameters below, as suggested by Supra:

|  |  |  |
| --- | --- | --- |
|  | **Parameters** | **Reasoning** |
|  | **Vintage** and **Age** | If a mature customer has a longer relationship with that product, there is a lesser likelihood of making changes to the status quo and churning |
|  | Percentage change between **account balances** | The trend in movement of balance and account activity is a good indicator of product engagement. An engaged customer is less likely to churn |
|  | **Days since Last Transaction** and **Vintage** | If a customer with a longer relationship has recent transactions on the account, then it denotes product engagement. An engaged customer is less likely to churn |
|  | **Age Group** and **Occupation** and **Vintage** | Customers of certain occupation at higher age groups and longer product relationship will be less likely to change status quo and churn |
|  | **Gender** and **Number of Dependents** and **Age** | All things being equal, the number of dependents dilutes the available credit limit. The gender and age of the customer is also expected to influence behaviour |
|  | **Net Worth** and **Occupation** | There is a correlation between the earning cap of different professions which reflect in the credit usage |
|  | (Tessy’s addition) All **balances**, **credits** and **Number of Dependents** | Divide all the balances and credit values by the number of dependents, which will give a metric similar to 5, but with more parameters to train the model and discover correlation |

## Handling missing data

**occupation**: 182 missing values. After a short review of if we can predict occupation given the current data, that idea was abandoned in favour of using “unknown” to fill the missing values

**current\_balance**: 715 missing values. After reviewing the possibility of imputing this value based on trend, it was decided that this will be imputed by taking average of other balance features for the same customerID. The advantage of this method is that the imputed value will be inline with the customer’s balance history. The alternate option is to group by gender and occupation, then using the median value of each group to fill the missing values

**previous\_month\_debit**: 887 missing values. This will be imputed by adding a random number to the available field current\_month\_debit. The random number will be determined by the average difference between the previous\_month\_debit and current\_mont\_debit. The alternate option is to group by gender and occupation, then using the median value of each group to fill the missing values

current\_month\_balance: 816 missing values. This will be imputed by taking average of other balance features for the same customerID. The advantage of this method is that the imputed value will be inline with the customer’s balance history. The alternate option is to group by gender and occupation, then using the median value of each group to fill the missing values

## Identifying parameters for feature engineering

The merits, demerits and the available data for the parameters listed above were reviewed by their suitability for model development. The sequence corresponds to the decisions taken on the suggestion listed above

1. Create buckets of vintage of 1000 days, create buckets by age. Two methods to be tested: equidistant grouping by span of 10 years, or equi-density grouping where each group has equal population. Scores will be assigned to each group, then the scores multiplied to derive the engineered parameter.
2. Calculate three percentage change of
   1. (current\_balance – previous\_period\_balance) / previous\_period\_balance
   2. The above formula will be applied to 6 consecutive period balance features.

These percentages will be used as engineered features

1. Calculate the ratio of (vintage / days\_since\_last\_transaction ) as an engineered feature
2. Will be reviewed later if needed
3. Will be reviewed later if needed
4. Take current balance, divide by dependent + 1 to get avg monthly balance
5. NW can be converted into ordinal number with low=1, med=2, high=3 is sufficient, ignore occupation

## Next steps

|  |  |  |
| --- | --- | --- |
| Phases | Task | Time of completion |
| 1 | Tessie will engineer the new features discussed and develop a pipeline | 12 April (Tuesday) |
| 2 | Using the Pipeline, build about 6 models. Tessie, Yalini and Supra will split the work between themselves and build 2 models each | 14 April (Thursday) |
| 3 | Compare the performance of Aleem’s models built on the features available in base dataset to the models built on engineered features | 15 April (Friday) |
| 4 | Performance tuning for candidate models | TBD |
| 5 | Write report and peer review | TBD |

Proposed models to develop and compare:

* SVM
* Ensemble model (e.g., Random Forest, boosting algorithms)
* KNN
* Decision Trees
* Logistic Regression
* Etc.