Technique report of the bank marketing campaign

Zheng Bin

10/08/2021

Contents

Introduction of the project

How to increase profits is the most important topic of every company. This report will show how a bank use machine learning methods to improve profit from direct marketing campaigns.

The problem is solved in two steps. The first step is to build regression models to predict the profit of each customer. The second step is to establish multi-label classification models to predict which of the 3 products each customer will buy.

Step 1: The regression models to predict profit

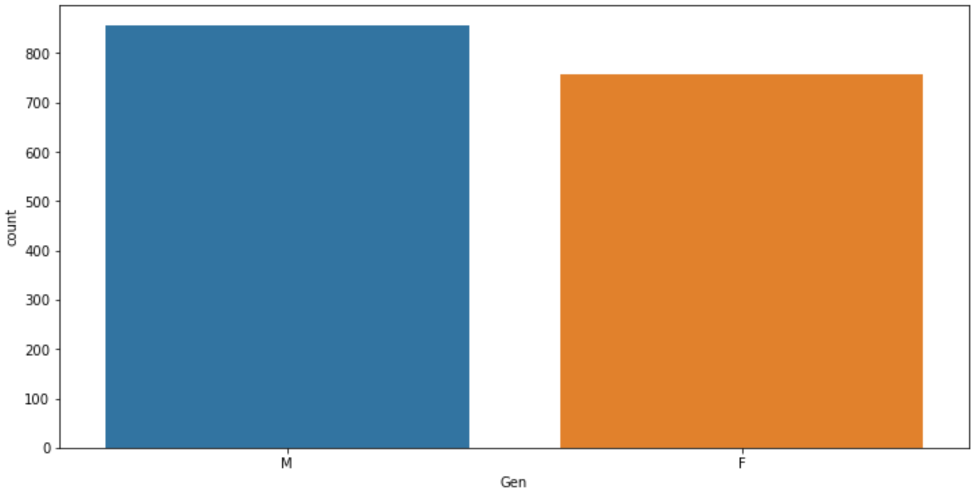
Data preprocessing

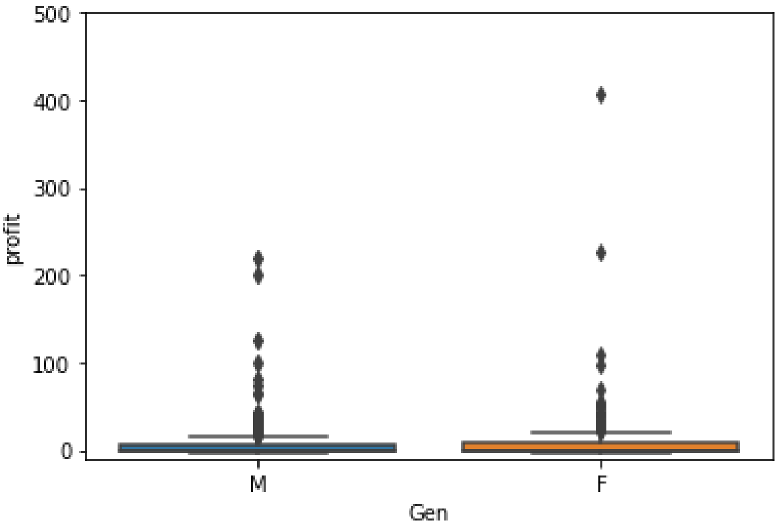
- Checking all the data

- For calculating profit of each customer in the training set have built 3 new features
  1. Number of products of each customer have bought
  2. Total revenue of each customer
  3. According to the equation $Profit = Revenue - Cost$ to calculate the profit of each customer.

- Merge all data with 'left join'
  For the sales and income data, only the profit variable that have been created in advance is retained as the target, because the test set does not contain the variables in the sales and income data.

- Check gender categorical variable

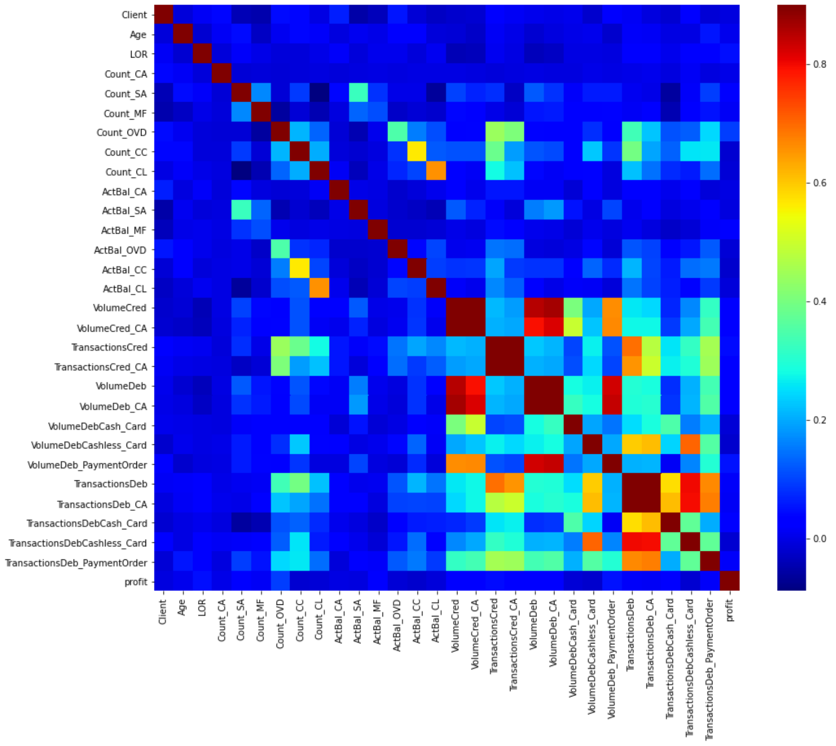




The histogram shows that the gender ratio of young women is relatively balanced, and the box diagram shows that different genders have no special impact on profits, but individual women contribute more than 400 to profits, far exceeding the men who contribute the most to profits. But this difference exists in real life, so the outlier is not dealt with.

- Missing values
  For numeric variables have fill the missing value with 0. Tracking columns have been created for each variable.
  For classification variables have been transformed with dummy coding. Tracking columns have been created for each variable.

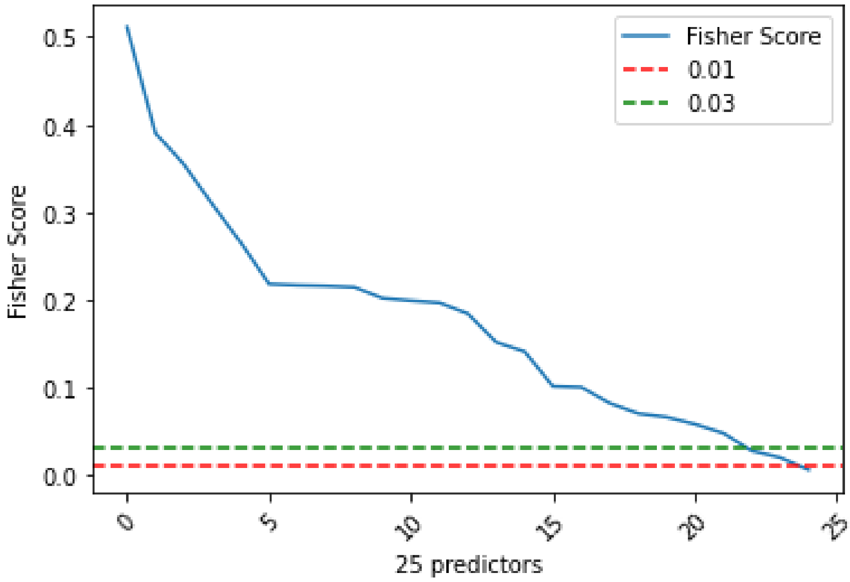
- Check the correlation between variables, and drop the high correlation variables



According to the high correlation between the variables, the tracking variable and the other four variables are removed, including VolumeCred, VolumeCred_CA, Count_CC and Count_CL.

- Partition data
  According to the missing value tracking variable profit_miss which has been built for the target splitting the data. (Proportion of test set is 0.4, proportion of train set is 0.6)

- Feature selection



According to the fisher score curve, the curve starts from the increase of the 15th variable, and the fisher score decreases relatively slowly. Hence the top 15 variables with the highest fisher score have been selected.

Modelling

- Resampling Methods
  1. Cross validation method: 5-Folds with shuffle
  2. Negative mean squared error
  3. Loss function: Negative mean squared error. (To minimize the MSE)

- Models and main Parameters
  1. Lasso
     - Robust Scaler (Lasso/Elastic Net model is sensitive to outliers)
     - alpha=0.0005
     - score: [Mean MSE: 18.8630, Std MSE: 9.4652]
  2. Elastic Net
     - Robust Scaler (Lasso/Elastic Net model is sensitive to outliers)
     - alpha=0.0005
     - l1_ratio=.9

        - score: [Mean MSE: 18.8630, Std MSE: 9.4652]

    3. Kernel Ridge
        - alpha=0.6
        - kernel='polynomial'
        - score: [Mean MSE: 35.7097, Std MSE: 23.0853]
    4. Gradient Boosting Regressor
        - Learning rate=0.05
        - score: [Mean MSE: 19.2038, Std MSE: 9.4060]
    5. XGB Regressor
        - Learning rate=0.05
        - score: [Mean MSE: 23.8923, Std MSE: 7.8613]
    6. LGBM Regressor
        - Learning rate=0.05
        - score: [Mean MSE: 21.0823, Std MSE: 8.3575]

- Model Selection
    1. The Elastic Net is the best model with the lowest Mean MSE
    2. Prediction with the Elastic Net model
    3. Acquire the top 120 customers with the highest profit, and the total profit contributed by these 120 customers is 1136 euro

Step 2: The Classification models to predict Products that customers could buy

Data preprocessing

- Merge all data with 'left join'

  For the sales and income data, the variables Sale_MF, Sale_CC and Sale_CL are retained as the multi-target, because the test set does not contain the other variables in the sales and income data.

Method of multi-label classification problem

1. Problem Transformation
   - Binary Relevance
     - Principle: This is the simplest technique, which basically treats each label as a separate single class classification problem.
     - Disadvantages: it doesn't consider labels correlation because it treats every target variable independently.
     - Performance: accuracy score is 0.1482758620689655

   - Classifier Chains
     - In this, the first classifier is trained just on the input data and then each next classifier is trained on the input space and all the previous classifiers in the chain.
     - Disadvantages: The lower accuracy is maybe due to the absence of label correlation
     - Performance: accuracy score is 0.1482758620689655

   - Label Powerset
     - Principle: It transform the problem into a multi-class problem with one multi-class classifier is trained on all unique label combinations found in the training data.
     - Disadvantages: As the training data increases, number of classes become more. Thus, increasing the model complexity, and would result in a lower accuracy.
     - Performance: accuracy score is 0.06896551724137931

2. Adapted Algorithm
   - Principle: Adapted algorithm is adapting the algorithm to directly perform multi-label classification, rather than transforming the problem into different subsets of problems.
   - Performance: accuracy score is 0.4517241379310345

3. Ensemble approaches (not applied here), normally have the better performance.

Model Selection

- The Adapted Algorithm is the best model with the highest accuracy score.
- Prediction with Adapted Algorithm

Acquire customers who can maximize the bank's profits

- Get all the customers would buy the products with their profit
- There are 104 of customers would buy the products and profit>0
- Get the other top 16 customers with highest profit who would not buy the products and profit>0
- Then concatenate the data of 104 and 16 customers
- The total profit of customers who could buy products is 711.29 euro

Summary

Through data processing and building machine learning model, after compared with the regression models have been built, the 5-fold Elastic Net model the best performance with the lowest mean MSE and relatively low Std MSE.

After establishing several Multilabel Classification models, and comparing the performance of the 4 different methods, it is found that the Adapted Algorithm model performs better, which with higher accuracy score.

Because profit forecasts and product purchase forecasts are carried out separately. Although there is no purchase volume for some products, there are still profit expectations. If the product and revenue can be predicted in advance in the same model, then the method of calculating profits may be more reasonable. To further improve the performance of the product prediction model, we can try to use Ensemble approaches that usually have better performance.

References

1. *https://scikit-learn.org/stable/modules/multiclass.html*
2. *https://www.youtube.com/watch?v=-EIfb6vFJzc&t=25s*
3. *http://scikit.ml/api/skmultilearn.html*
4. *https://stratos-initiative.org/sites/default/files/HEC16-Heinze-TG2-Varsel.pdf*
5. *https://onlinelibrary.wiley.com/doi/pdf/10.1111/tri.12895*
6. *https://www.youtube.com/watch?v=bN0OU7jeObI*
7. *https://scikit-learn.org/stable/modules/model_evaluation.html#scoring-parameter*
8. *https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html*