BA 820: Final Project Paper

Team 2: Yulong Gong, Muyan Xie, Yangyang Zhou, Yichi Zhang

I. Business problem

Nowadays, more and more companies realize the value of data, especially those collected from their business cycle. Not only can the data help them design products according to the natural clusters identified from their business data, but also it could help them identify certain customer behavior patterns, such as churn/ no-churn. The use of data mining techniques for predicting customer churn is new in the electronic banking context. Similarly, actual reasons behind the customer's churn decision might affect the bank's energy to explore new products or how to allocate the marketing resources

Through this project, we hope to help the business manager understand their customer stratifications and update their product features. Furthermore, we want to use the information we have to identify the characteristics of attrited customers so that the company may come up with marketing strategies to maintain those customers who may end the business relationship with them.

II. The dataset

The dataset was from Kaggle: https://www.kaggle.com/sakshigoyal7/credit-card-customers

Target features to focus on in this project are:

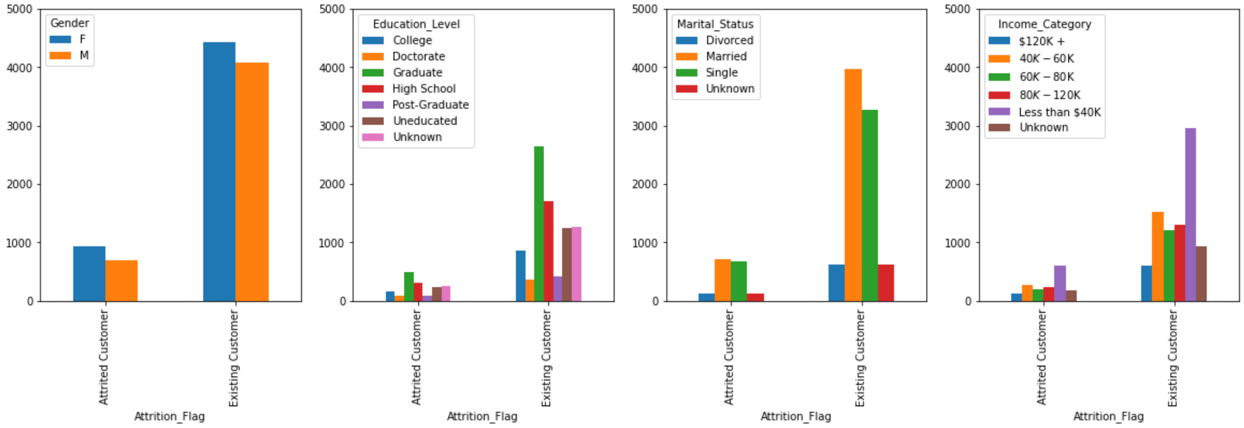
- Attrition_Flag
- Card_Category

Relevant features contain but is not limited to:

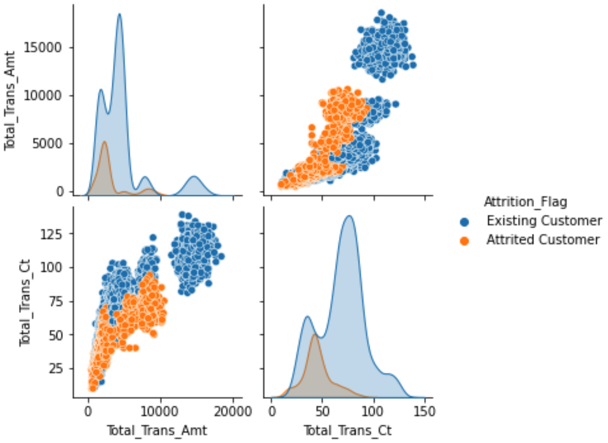
- Gender
- Credit_Limit
- Average_open_to_buy
- Total_ct_chng_q4_q1
- Avg_utilization_ratio

III. Exploratory Data Analysis

We load the file from Kaggle into a python environment and then proceed to prepare a dataframe for our analysis by standardizing our numeric columns and dummifing object columns.



- To consider potential data sampling issues, we checked the data distribution in terms of gender. Surprisingly, as shown in the first figure, there are more female records than male records. While the figures indeed show some disparity in numbers between churned and existing customers, the distribution of each category is very similar. Also to support the claims of the previous figure, we see again that age does not play a major factor in churned and existing customers. The figures above show that each category cannot be used alone as a factor to decide customer churn. A married female with higher income and education is a different demographic than a single female with medium income and education. As such, we will need to weigh in all these features to build our model.



- Another view to look at the data is, as shown in the chart on the left-hand side, that existing customers tend to have more transactions than attrited customers. Existing customers tend to spend more than attrited customers. In other words, the more frequently you use your credit card, the more active you will be and the more likely you will continue the business relationship with a company, in this case, the bank.

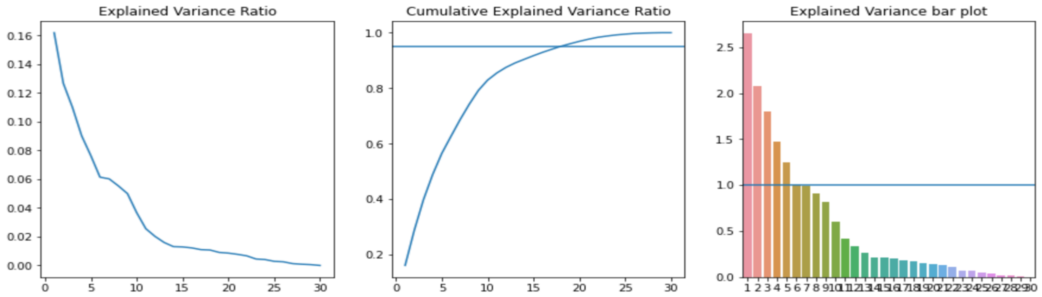
IV. Methods:

To solve the business problem we proposed, we decided to utilize unsupervised machine learning to help us reduce feature dimensions and identify natural clusters in the dataset. Furthermore, we want to combine the results we have with supervised machine learning so that with a new observation we could make accurate predictions in terms of product recommendations and marketing strategies.

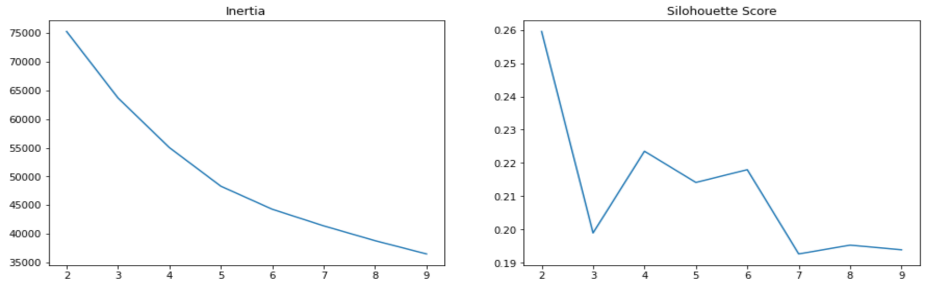
Before training our models, we did some data preprocessing. First, we separate the numeric and categorical data. Then, we standardized our numeric data so that they are on the same scale. We also dummified our categorical data to give them more predictive power. Last, we combine the standardized numeric data and dummified categorical data together to get our final data frame.

1. Card type prediction:
   a. PCA + KMeans model

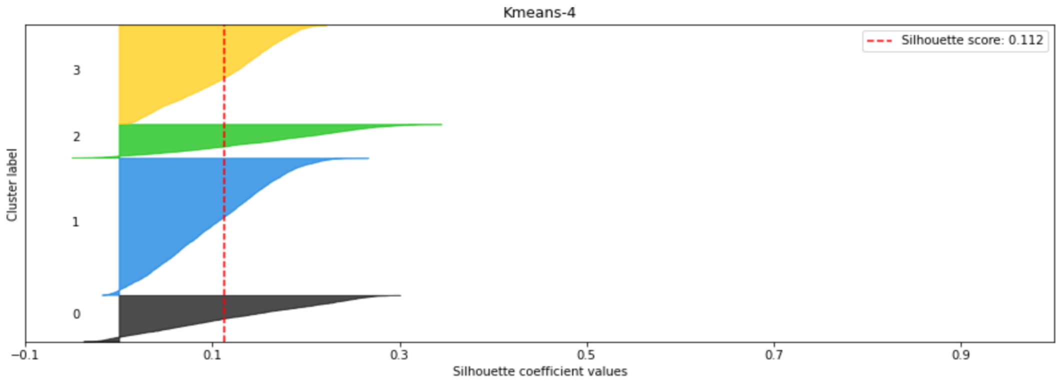
      We’ve tried to use Hierarchical clustering, but we have more than 10,000 observations and it takes too long to run a single dendrogram, and the result is not so satisfying. Therefore, we decided to use KMeans. Since we have 21 features for each observation, we want to reduce dimensions to get a more accurate result. Firstly, we chose to use PCA. From those 3 plots shown below, we decided to keep 5 pcs for further clustering because 5 pcs explained about 80% of the variance.



Then, we tried to find the proper k value for KMeans. Our dataset is about credit card customers, and there might be some business constraints. We were thinking k should be in between 2 and 4, because there are 2 kinds of customers (churn/no-churn), and 4 kinds of card types (blue, silver, gold, platinum). Moreover, as the graph is shown below, k=4 is also a good choice because it has low inertia and a high silhouette score.

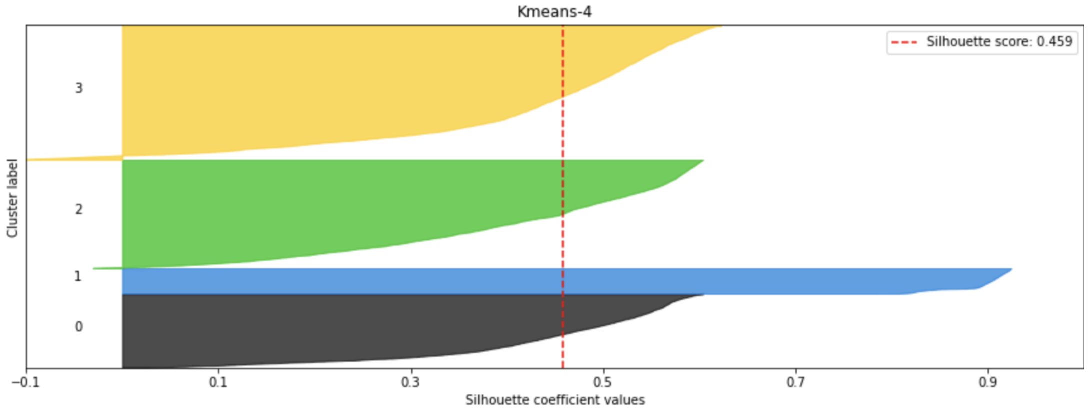


After deciding the k values, we fit the KMeans model and get a 0.112 silhouette score, and from the graph, we can see that although there are some misclustering exist in group 0,1,2, but the model performs well overall. We have relatively the same size of data in each group as in each card type.



b. UMAP + KMeans model

We tried UMAP as another dimension reduction method. After compressing our dataset into a 2-dimension space, we used KMeans to separate the observations into 4 groups. As the graph is shown below, the UMAP method has a higher silhouette score than the PCA method. However, when we profile the cluster results, the characteristics of each group are not distinct, and we cannot get useful information about the customers in each group.



2. Churn/ no Churn:

Since one of our goals is to predict whether customers would churn credit cards, we decided to try out two analytical techniques - unsupervised machine learning and PCA with supervised machine learning – and examine which one yields better results.

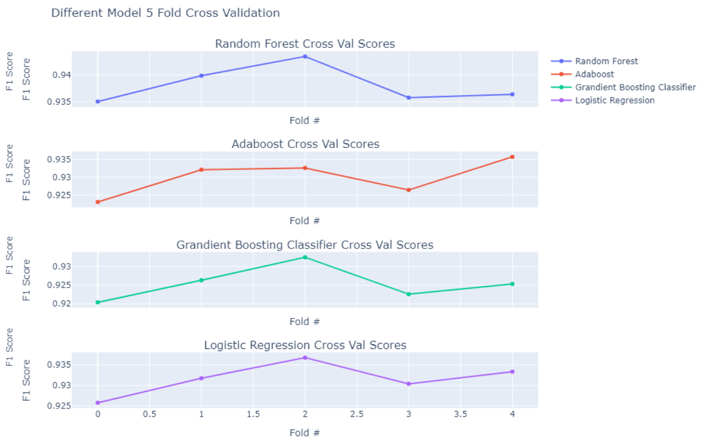
i. Original data with unsupervised models

We tried to use the unsupervised models with the original data. Since we have two types of customers, we want to make 2 group clusters to see the natural patterns of our clients. Then we used the KMeans model, and we found that only three features are significantly different: the credit limit, gender, average open to buy. We can't distinguish groups well from these features for the patterns are not obvious.

ii. Original data with supervised models

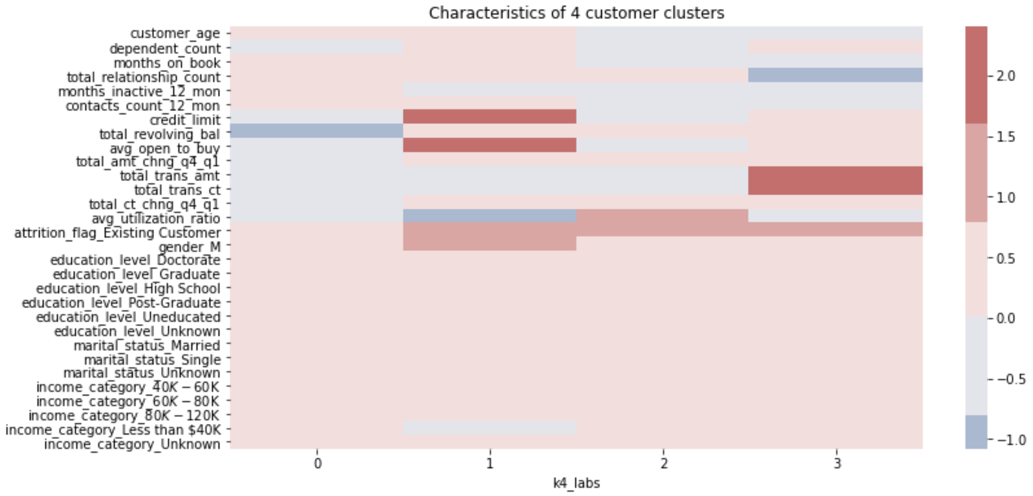
We tried to use the supervised models. First, we used the original data with F1 scores to balance accuracy and recall. Then, we found that the f1 score of the logistic regression model is 0.9 and for the random forest model is 0.88. And then from the logistic regression model, we also observed several of the most important features. We found that total revolving balance, transaction amount, and transaction count have a significant impact on whether a customer is existing or not.

iii. PCA with supervised models



To improve model accuracy, we tried to combine supervised and unsupervised models, using the data after PCA to import the supervised model. We used the above PCA data (5 pcs) for further modeling. The chart, as shown on the left-hand side, shows the f1 scores of these four models after five cross-validations. Compared to Logistic Regression, Adaboost, and Gradient Boost classifier models, Random Forest has the highest f1 score of 0.94. Therefore, after tuning the PCA, there is a significant change to the f1 score and accuracy of the model.

V. Findings



Cluster 0: There are 3167 observations in this cluster. They have a relatively low total revolving balance and relatively low transaction amount. Their credit limit is kind of low and they are less likely to spend huge amounts of money.

Cluster 1: 1486 observations fall in this cluster. They have a high credit limit and high open-to-buy amounts. Their utilization rate is relatively low, which might imply they have multiple credit cards. In this cluster, the majority of customers are men.

Cluster 2: This cluster includes 4398 observations. They have a high utilization ratio and more likely to be existing customers. They are also less likely to spend huge amounts of money and make large amounts of transactions.

Cluster 3: This group contains 1076 observations. Customers in this group have high transaction amounts and transaction counts. They have relatively low relationship counts. If the company wants to reduce the churn rate in this group, introducing more products to those customers would be helpful.

VI. Conclusion & Recommendations

Conclusion:

For unsupervised machine learning, if we simply perform k-means clustering after PCA. The results did not come out well for churn/no-churn. And silhouette score will stay low around 0.1. And DBSCAN does not have a good result either. After we imported TSNE and UMAP methods, there are still no good outcomes at attrition classification.

However, UMAP stands out when we conduct this method with k equal to 4 for card types, getting a silhouette score of 0.45. But there are no obvious patterns for groups with different labels. The result is not convincing overall.

Then, we try to use supervised methods to explore the dataset with the customer churn/no_churn as the target variable. PCA helps improve the overall performance of all models we use. After PCA, we acquired the best F1 score of 0.94 for the Random Forest model.

Finally, in the clustering process, two variables attrition and card type both did not perform well. But we get a decent number for card types in the prediction process. We would like to say that the attrition pattern is hard to describe according to the customers' features we have in this dataset. And it is not balanced because more than 80 percent of one kind. However, we still would like to say card type is a significant indicator for the bank to perform user filtering.

Recommendations:

- Banks could make card types more identifiable. We hope more patterns could be found in a certain type of card. And Platinum card users do not show a significant difference with other types. For example, banks could approve higher limits for those active customers.
- In our clustering practice, the gender distribution is almost even, however, gender becomes one of the features which is different for the two clusters case. The bank should evaluate its data integrity to check whether gender could make a difference. If this is true, we would recommend applying different marketing strategies for males and females.
- Cooperating with other businesses to stimulate consumption which may be helpful in terms of maintaining those customers who may churn.
- According to the variable importance, customers having more relations with the bank tend to stay. We recommend the business expand business with existing customers.