Name: Viet Duong

DSC 383W Data Science Capstone

Data Analysis Report

Paychex Time-series Analysis Project

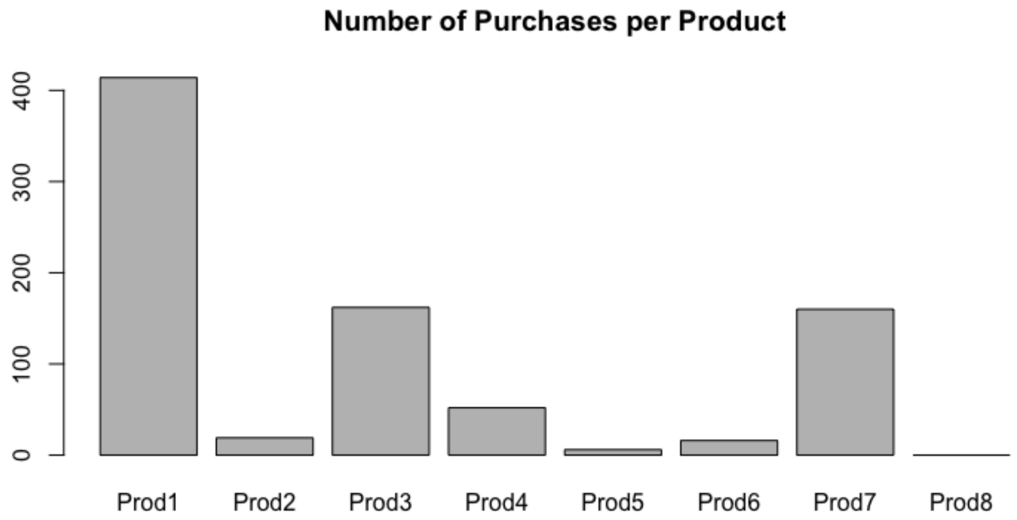
Introduction

Due to advancement in computer science and software engineering, a growing number of data analysis techniques is being integrated into the strategic planning and decision-making process of corporate firms. Instead of hiring market analysts to "beat" the market using their qualitative market knowledge and economical intuitions, the focus of business strategists has been shifting towards more quantitative approaches in order to take advantage of the large databases amassed by collecting data on transaction records and other customer interactions. Paychex, being one of the largest American human resource service provider supporting over one million employees, would benefit greatly from collaborating with data specialists to learn resourceful information from their own databases. In order to sustain sales growth and target the appropriate demographics for marketing and reach-out programs, being able to predict customers clients at certain points in time is a great advantage. Therefore, our group is entrusted by Paychex to provide some insight into their database using time-series analysis and produce results that support their business-related decisions. This report aims to describe our data preparation process and exploratory analysis on the Paychex data, as well as proposing our modelling approaches and plans towards the end of the project.

Data Collection and Preparation

Paychex has provided us with a data package consisting of 10 tables. Eight of the tables record the probability scores corresponding to the likelihood of purchase of their 8 individual products on a monthly basis from January 2016 to December 2016 for 25,000 customers with unique ID numbers. One table shows the probability that a customer became ineligible to purchase any product from Paychex after some point in time within the year 2016. And one table records the dates of activities of 7500 customers who either purchased any product or became ineligible to purchase from Paychex during the January 2016 – September 2017 period.

Since the core objective of the project is to learn the customers' behaviors from the trends in probability scores, it is important to highlight the probability instances where any kind of activities occur in order to evaluate the meaning of such values. We also observed that out of 25,000 customers who are included in the probability tables, only 7,500 have their activities recorded. Hence a subset of customers with some activity out of 25,000 was extracted from the each of the 8 probability of purchase tables by matching their customer IDs and product number with their respective occurrences in the activity record, as well as computing the associated months of purchases in an additional column. The sample sizes of the new 8 tables are shown below:



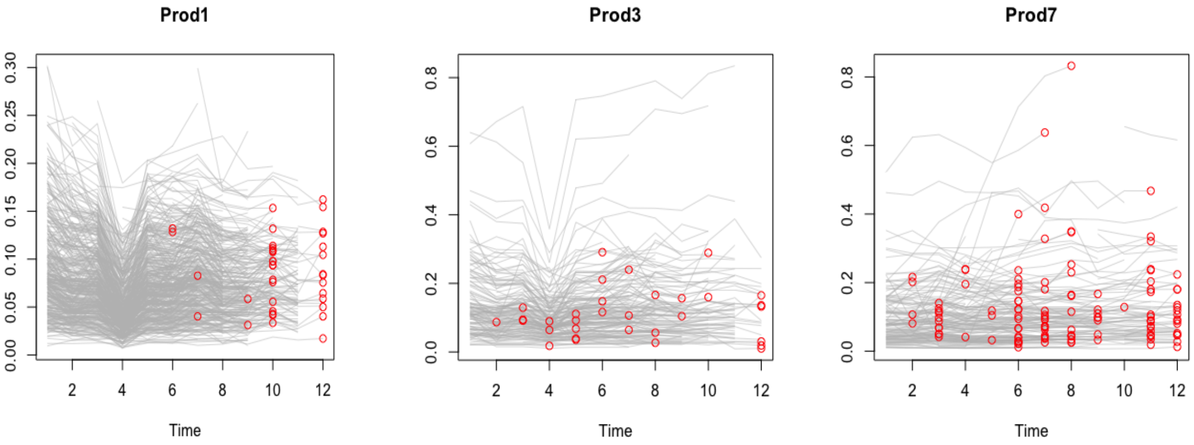
We repeated a similar process with the probability table of the likelihood that a customer became ineligible to purchase from Paychex and create a dataset of 1,276 customers with their probabilities of becoming ineligible for the January 2016 – December 2016 period as well as the months of such events.

Last but not least, the activity data during January 2017 – September 2017 is going to be our test dataset to examine how accurately and efficiently our models model the probability trends and detect the respective customer actions.

Exploratory Analysis

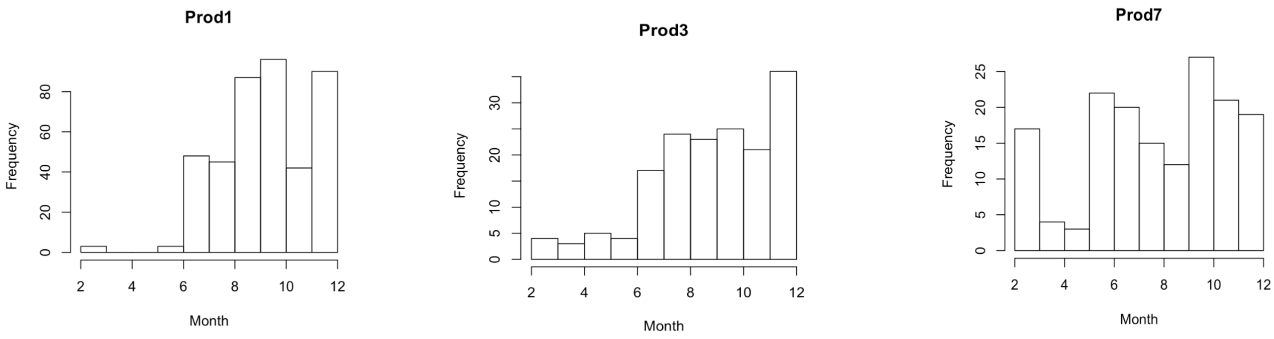
For our exploratory analysis, only the samples for Product 1, Product 3 and Product 7 were used since they have sufficiently large sample sizes (greater than 100), which is useful for us to visually observe and assess the time-series probability trends and the distributions of the probabilities at the months of purchases.

The first step is to plot the time-series probability trends for each sample, with the highlighted (in red dots) probabilities of the months of activity. Following the objective of the project, that is predicting client actions using the trends in probability scores, it is useful to model the probability trends exhibited in the data and use the information on the points of activities to learn if such activities are likely to occur.

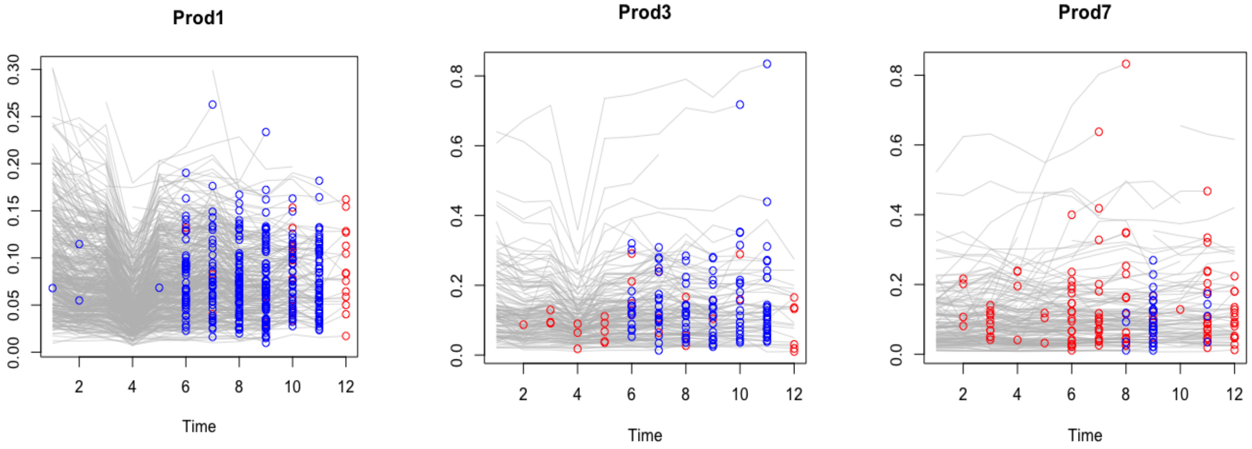


Observing our plots, the probability trends are very noisy, which is one of the characteristic of time-series data. However, the real issue is the amount of data available, that is there are at most 12 data points per customer. This means the trends are difficult to model, and the prediction accuracy for subsequent periods would be consequently poor. Also, the probabilities at the months of purchase are widely spread out. Hence, using one probability value or a small fixed confidence interval to predict client actions is not possible.

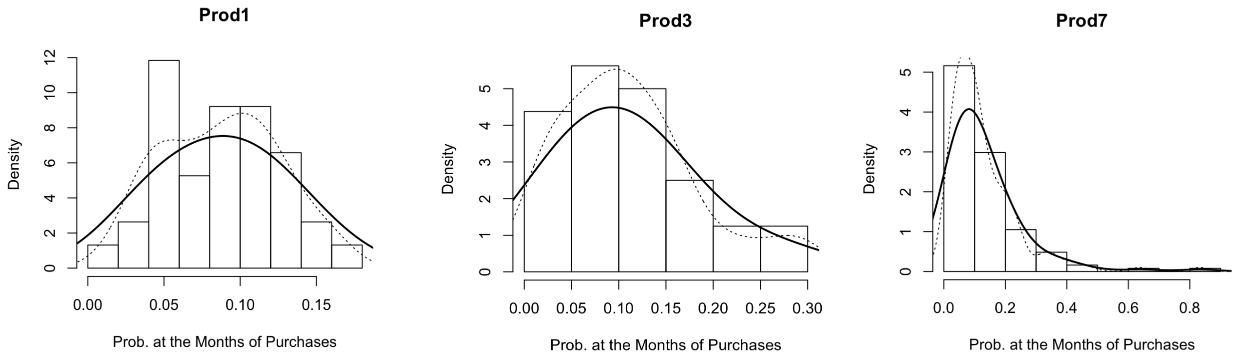
Despite the volatility of the data, there are some client actions supported by the probability trends. The most visible pattern is the lack of purchasing activities of Product 1 within the first four months of 2016. The absence of activity during this period can be predicted as the probabilities for almost all the customers who purchased Product 1 follow a downward trend, reaching the bottom on April. However, this is not supported by the graph of Product 3. Despite exhibiting similar probability trends to Product 1 during this four-month period, the purchasing activity for Product 3 in April is not significantly less than the other months. Also, although there was significantly less activity in April for Product 7, there is no significant drop in probability values for most of the customers. In fact, the majority of purchases happened when the probability trends were fluctuating within a very small range of values, which might be considered as flat. Therefore, the surge of activity for all the products within the last six months of the 2016 seemed impossible to be explained by the probability trends. Below are the histograms showing the activities for each month during the January 2016 – December 2016 period:



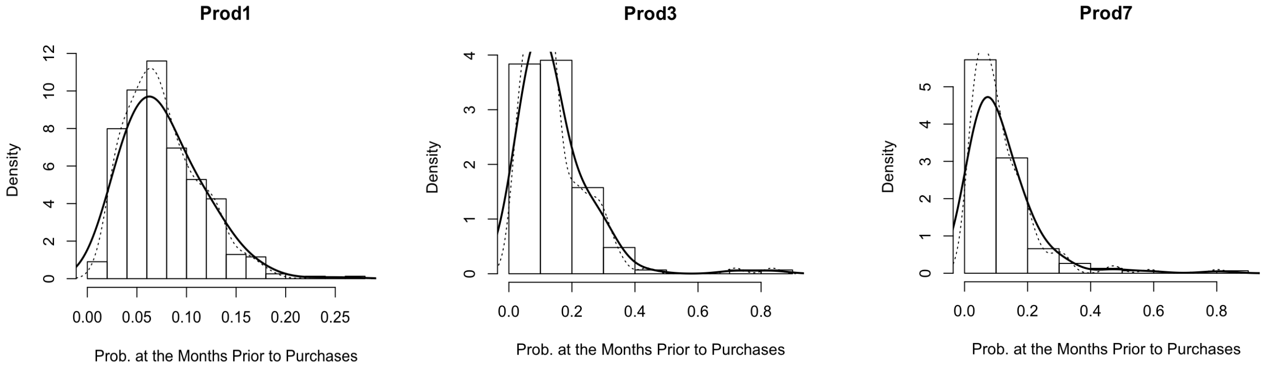
Another problem that we encountered when plotting the data is that there are many missing probability values at the months where the products are purchased, as demonstrated by the blue points in the graphs below:



Hence, we attempted to plot the distributions of the probabilities at the months of purchase in order to evaluate their relationships. The probability values at the months of purchase were sorted in ascending order. Then they were divided into 10 bins of equal width, computed by dividing the maximum probability value by 10. The height of each bin corresponds to the density of a range of probability values. The histograms for the densities of purchase probabilities for Product 1, Product 3 and Product 7 are shown below:



To observe the characteristics of the distributions, we plotted the probability density function (PDF) of purchase probabilities (the dotted curve) for each histogram. A smoothed version of each PDF is also plotted, dividing the slope at every point on the PDF in half (reducing the difference of densities between each pair of consecutive probability values by half). The same process is repeated to visualize the distributions for the probabilities at the months prior to purchase for the three products.

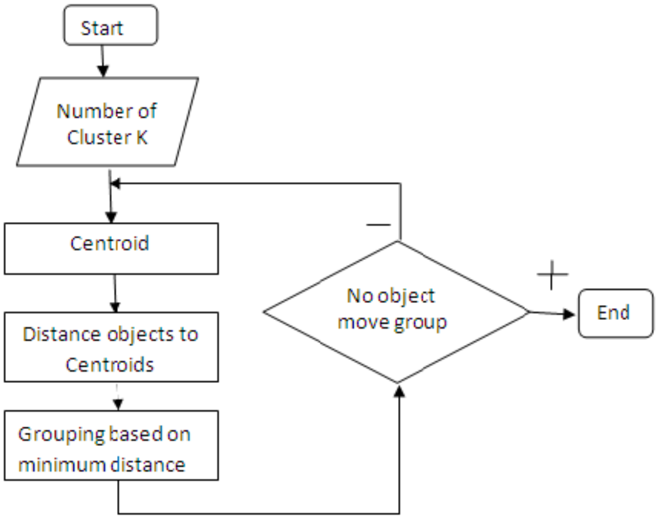


Out of the three products, the distribution of probabilities at the months of purchases of Product 1 resembles a normal distribution the most while all the other density curves are right-skewed. This means that the majority of products are purchased when the probabilities are very low. In addition, all the distributions of probabilities at the months prior to purchases are right-skewed as well. This strongly supports our observation that the probability trends are almost flat when purchases are made, since if the purchases are made at higher probabilities, the curves should be skewed in a more positive direction. Only the distributions on Product 1 suggests that the probabilities might increase at the months of purchases, but having only 38 probability values at the months of purchases of Product 1 is not enough to tell us anything concrete.

In summary, initially, we proposed to predict the probability trends using time-series regression models then use a fixed threshold such as the mean of probabilities of purchases or becoming ineligible to detect client actions. However, this approach is not supported by our exploratory analysis on the datasets. After presenting our preliminary results, as well as consulting with the Paychex team and Professor Anand, we decided to use clustering techniques to distinguish groups of Paychex customers with similar characteristics in their corresponding probability trends. Then, we will label the activity levels for the resulting clusters of customers and validate the clustering performance based on the activity data. Also, we learned from Paychex that the activities on Product 1 and Product 3 were influenced by taxation in April, which explains the extreme dip in probability trends. Since we have no data or control regarding this matter, we will examine Product 7 individually to eliminate externalities.

Methodology

First of all, let us give a brief overview of clustering analysis. As defined by Jiawei Han, Micheline Kamber and Jian Pei (2006), clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, which consists of objects that are characteristically similar to one another, yet dissimilar to objects in other clusters. The clustering technique chosen for our analysis is K-means. It is a clustering method which separates the data into a predetermined number of clusters by finding the optimal set of center points, also known as centroids so that the geometric distances between the data objects within each cluster and its center point are minimized. The overall process follows the outline in the figure, and we will explain in detail the data manipulation steps that will result in the final clustering of the Paychex customers.

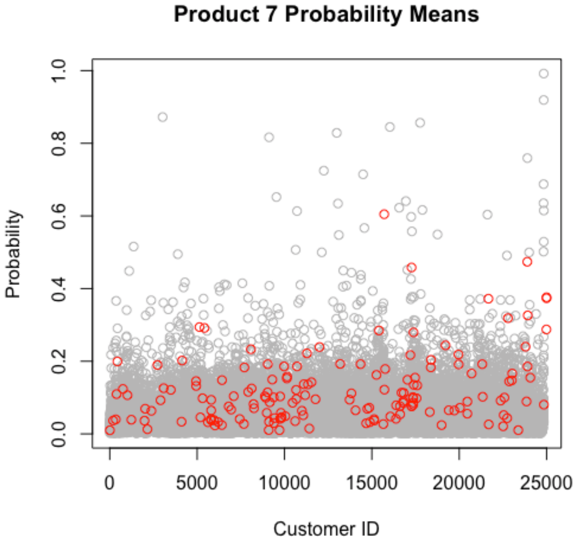

Start
Number of Cluster K
Centroid
Distance objects to Centroids
Grouping based on minimum distance
No object move group
—
+
End

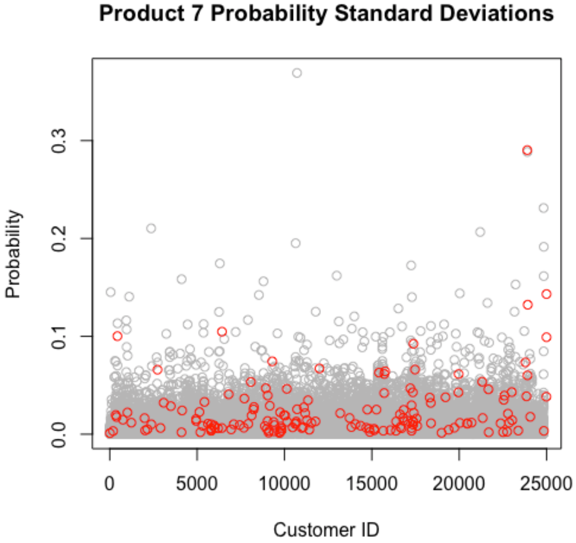
Since the K-means algorithm operates with a given number of clusters, it is vital to decide the appropriate. As we were striving to identify groups of customers who have similar levels of purchasing activities or likelihoods of ending their business with Paychex, it is reasonable to have 3 clusters of high, medium and low probabilities of such events.

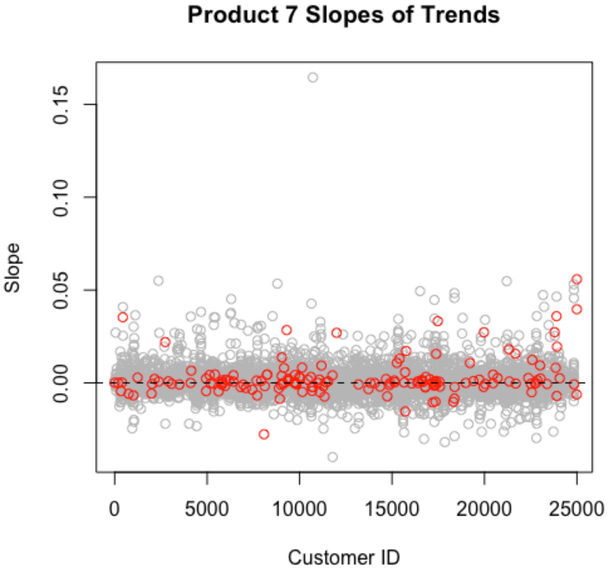
Also, in order to generate good clustering, feature engineering is a very important task. We expected “good customers” might exhibit some distinguishing characteristics such as having high mean purchase probability, less volatile probability trends, positive slope of the linear regression line and less fluctuations between consecutive months. Therefore, these four features were selected to represent the above characteristics:

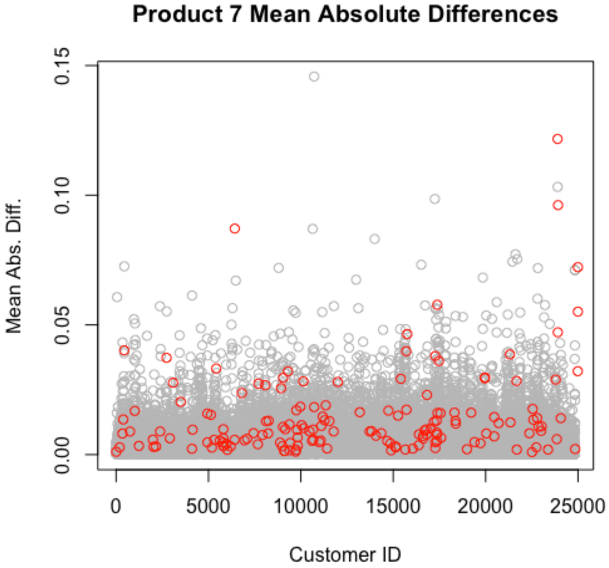
- Slope of Probability Trend: the slope of the linear regression line fitted on the probability trend for each customer.
- Probability Mean: the average of the probabilities over the 12-month period in 2016 for each customer.
- Probability Standard deviation: the standard deviation the probabilities over the 12-month period in 2016 for each customer.
- Mean Absolute Difference: the mean of the absolute differences between every two consecutive months in 2016 for each customer.

The scatter plots for the 4 features computed from the probability table for Product 7 were shown below, with the values for customers who purchased Product 7 highlighted in red dots.


Product 7 Probability Means


Product 7 Probability Standard Deviations


Product 7 Slopes of Trends


Product 7 Mean Absolute Differences

Observing the densities of the red dots in several regions, we could see that the customers who purchases the Product 7 tended to have higher probability means and low standard deviations. Also, the slopes of their regression lines, as illustrated by the corresponding plot, could be either positive or negative, which suggests that the tendency to go up or down in probability does not affect one's likelihood to buy a product. The plot for mean absolute difference also showed no noticeable trait of a promising customer. Hence, the latter two features might be dropped off from our set of features. Further analysis could be drawn when we look at the clustering results.

With these four numeric features computed, each tuple of such four values represents a customer as a point in the four-dimensional space. Hence, the dissimilarity in characteristics between the customers can be parameterized by the geometric distances between their corresponding data points. In our K-means clustering algorithm, we chose to use Euclidean distance, which is given by the following function:

$$distance(customer_i, customer_j) = \sqrt{\sum_{x \in features} (x_i - x_j)^2}$$

The next step is finding a set of 3 cluster centroids to initialize the clustering algorithm. Typically for K-means, the initial cluster centroids are randomly selected. However, we learned from our exploratory analysis that the probability trends are tightly grouped together. So, if our choices of initial centroids are too close to one another, the clustering performance will be negatively influenced because the resulting clusters are more likely to overlap. Therefore, we used K-means++ initialization method, proposed in 2007 by David Arthur and Sergei Vassilvitskii, to ensure an appropriate spread of cluster centroids. K-means++ accomplishes this by picking subsequent centroids with probabilities proportional to their distances to the previously picked centroids, which means the set of centroids far away from each other is more likely to be chosen.

After the seeding of initial centroids are completed, the other data points are assigned the nearest centroids based on Euclidean distances. Then a new set of centroids were re-calculated by computing the mean values for the four features in each cluster, and the other data points were re-assigned to the new cluster centroids. This process is repeated until the clusters are stable, meaning the new assignment cannot improve the last assignment, which results in the final clusters.

Specifically, for this project, we ran the K-means algorithm with 100 different K-means++ initializations, averaging the clustering results in order to partition 25,000 customers into 3 clusters.

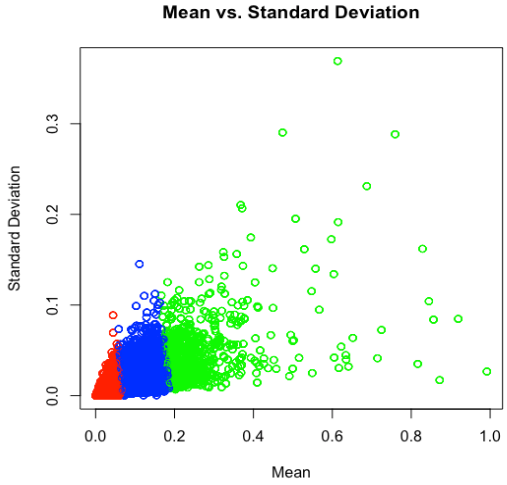
After performing clustering, the probability of actual purchase for each cluster was computed to measure how well the model partitions the probability data for Product 7 within the 12-month period in 2016. Let PP denote purchase “Purchase Probability” and N denote the “Total Number of Customers in Cluster”, the probability of purchase for each cluster of customers is given by:

$$PP(cluster\ X) = \frac{Number\ of\ Actual\ Purchaser\ in\ X}{N(cluster\ X)}$$

The clustering results was then projected into 2017 and probability of Product 7 purchase for each cluster was computed using the activity data during January 2017 - September 2017 (the test dataset) in order to validate the ability of our clustering method to predict the correct levels of customer activity in the future. The same techniques would be applied to train and test the model for the probabilities that a customer would become eligible to purchase from Paychex.

Results

Our clusters for Product 7 were nicely separated and exhibit distinguishable characteristics, especially regarding probability mean and standard deviation. This could be illustrated by our plot of probability means against probability standard deviations:



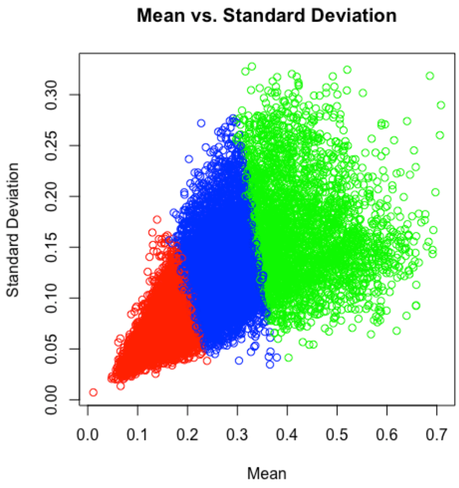
The high-value group of customers, denoted in green, had 751 instances and tended to have a higher mean probability. The low-value group, denoted in red, had 18273 instances and tended to have a lower mean probability. The medium-value, denoted in blue, had 6034 instances and the mean probability was somewhere in between. Also, the PP statistics for the clusters strongly supported that higher-value customers have higher probability of Product 7 purchase.

- High-value customers (green): PP = 4.66%
- Medium-value customers (blue): PP = 1.17%
- Low-value customers (red): PP = 0.03%

In addition, the probability of purchase for the high-value group of customers was significantly greater than the low-value group. Hence our clustering algorithm performed very well in separating these two groups of customers. Furthermore, the validity of our clustering algorithm was also strengthened when we applied the statistics to the test set (the activity table from January 2017 to September 2017):

- High-value customers: PP = 1.07%
- Medium-value customers: PP = 0.49%
- Low-value customers: PP = 0.18%

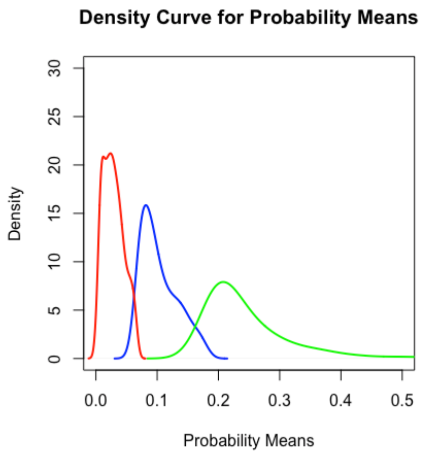
We also applied the same procedures to the probability table which predicted the likelihood of customers leaving Paychex, and the result was similar for those customers (LP):

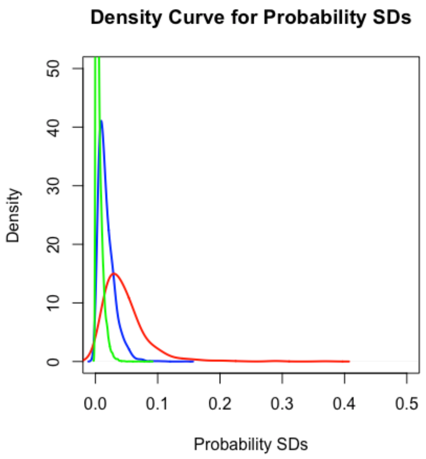


- Likely-leaving customers (green): N = 3240, LP = 13.52%
- Less-likely-leaving customers (blue): N = 9996, LP = 5.72%
- Least-likely-leaving customers (red): N = 11767, LP = 2.26%

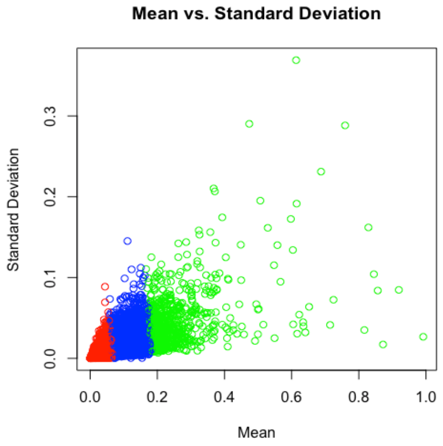
These numbers showed that our clustering algorithm performed just as admirably in predicting leaving behavior of the Paychex customers.

As discussed in the feature engineering section, the mean and standard deviation of those probabilities might be the most prominent features of our clustering algorithm. This can be observed as we plotted the density curves of probability means and standard deviations for different clusters of customers pertaining the Product 7.





Each cluster was separated nicely against these two features, with high mean probabilities and low standard deviations for the high-value cluster (green curve). Conversely, the low-value cluster (red curve) has low mean and high standard deviation. Recognizing the immense impact of the mean and standard deviation, we ran the clustering algorithm with only these two "key features", and the result from the previous four-feature clustering was mostly preserved:



- High-value customers (green): N = 780, PP = 0.0448717948717949
- Medium-value customers (blue): N = 6034, PP = 0.0117666556181637
- Low-value customers (red): N = 18186, PP = 0.00296931705707687

With only these two "key features", the cluster quality was not significantly deteriorated. This finding suggested that we could just use these two features instead of all four stated ones, which would save a lot of resources and runtime in future applications with larger-scale dataset.

Conclusions

In conclusion, we have successfully devised a clustering algorithm to group customers into partitions that exhibit their value to Paychex and likelihood of becoming ineligible to purchase products from the firm.

Regarding the probability trends for the customers, we found out that customers with higher mean probability score and less volatile probability trend for a certain product would be more likely to purchase such product in the future.

The likelihood of a customer becoming ineligible to purchase products from Paychex could also be predicted by our clustering technique, where the group that is more likely to become ineligible also has higher mean probability of such event, but having lower probability standard deviation.

Based on the density curves for 4 different features (only those for mean and standard deviation are shown), the mean and standard deviation of the available monthly probabilities are the most significant features. The slope of probability trend and mean absolute difference only improve the clustering result to a minimal extent. So, Paychex's data analysts can just deploy these two "key features" to work with larger-scale datasets.

In addition to the probability trends, the size of each cluster might also provide Paychex's managers with useful information to distribute their resources, since higher value group of customers tends to be smaller.

Suggestions for Further Work

With our findings on the impact of mean and standard deviation of customer's probability scores, significant regressors could be computed in order to construct predictive time-series models to generate future probability trends as well as updating the clusters of customers accordingly. These regressors can improve the probability of purchase statistics, as well as narrowing down the size of the high-value group of customers, which would promise a higher return when devoting the firm's resources to this group.

Also, the relationships between the probability of purchase data and the likelihood of ineligibility data might be further explored, so that more useful information could be extracted to detect non-benefiting customers for the firm.

Contribution

Since we were all getting familiar the programming language as well as the dataset throughout the process, we felt that it would be better if we all did everything together to fully understand the project instead of separating the work. Throughout the process, we did the everything together (each member contributed ideas, and syntax in R, and there would be one person who scripted out the code). Also, Viet worked as our main communicator via email, while Phu and Puching were the main presenters during in-class presentation. We took this as a great learning experience to grow our analytical skill and communication skill, as well as to learn from the sponsor's expertise. We also wanted to acknowledge professor Ajay Anand, who has kindly guided us throughout the process.

Reference

Jiawei Han, Michelin Kamber, Jian Pei. “Data Mining: Concepts and Techniques.” 3rd edition. Morgan Kauffman Press, 2011.

Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.