CSE145 Homework 2

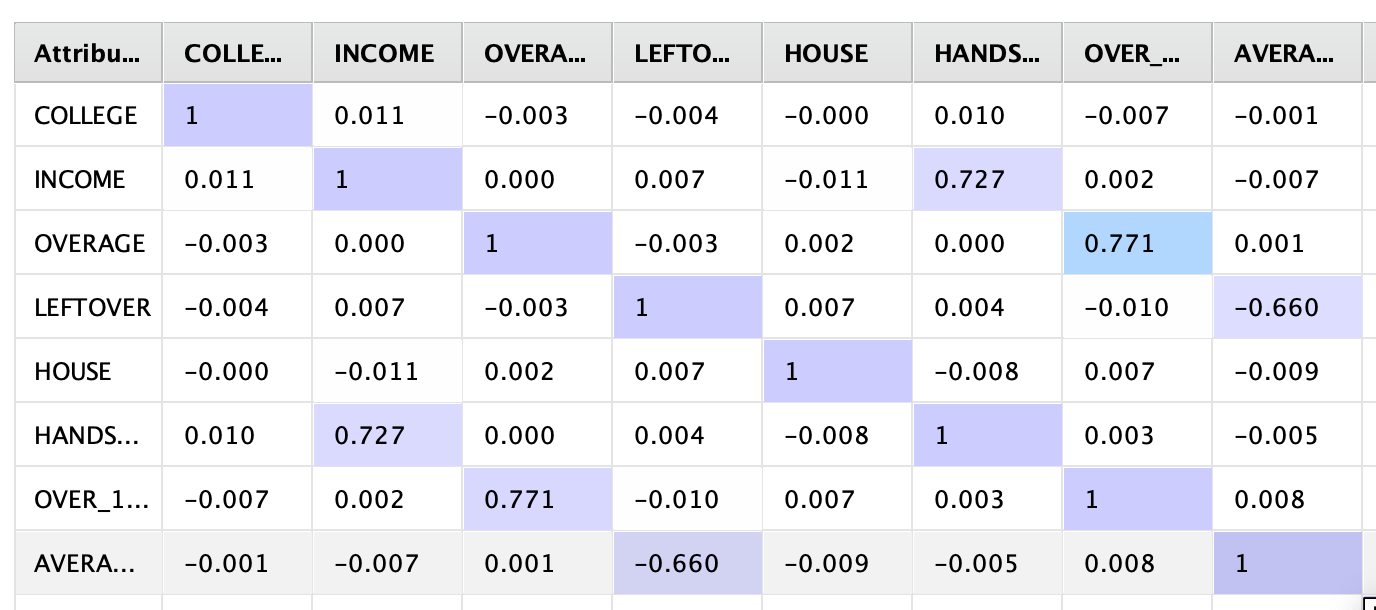
Eagle Huang

Customer Churn

1. **Data Understanding**: Describe and summarize the data. This is very useful as a preliminary step to capture basic data property. Distribution analysis, statistical exploration, correlation analysis, suitable transformation of variables and elimination of redundant variables, management of missing values. Produce at least two visualizations of the data that are meaningful to the report. Explain (in 1-2 sentence(s) for each figure) why do you think them are meaningful

The data contains information about 20,000 examples over 12 variables describing features of customers of a mobile phone provider. The attributes describe the customer’s social status and their usage of the phone service as well as whether or not they chose to churn. The data consists of no missing values but does have some redundant attributes that should be eliminated as they would result in unreliable and unmeaningful information. The most notable attribute that should be eliminated is the “handset\_price” attribute because the information that it tells us is not useful in predicting the targeted attribute or helping us decide on any future actions to take to improve consumer retention. The information that it tells us also overlaps with the income attribute.

Further analysis of the attributes through a correlation matrix tells us that some attributes are closely correlated.



Looking at the above correlation matrix, it is evident that over-15-min-call and overage are closely related which may be useful as consumers that churned who fall into this category could be persuaded to stay through special discount for long calls.

In terms of data preprocessing, the ordinal attributes should be transformed into numerical data so that they can be utilized during further processing. Attributes such as “reported-satisfaction” are ordinal, each with 5 distinct values ranging from very unsatisfied to very satisfied. These values will first be transformed from nominal to numerical, transforming into the range -2 to 2. The binary attributes like Leave or College will also be transformed into numerical binary. In addition, attributes with dollars as the measurement should be normalized as the scale is significantly greater than other attributes.

Finally, the distribution of some of the attributes, such as income and call duration, are skewed positively skewed (skewed left) while nominal attributes like satisfaction are bimodal. This suggests possible clusters within the dataset which would become clearer after clustering.

Graphical user interface, application, table

Description automatically generated

(Bimodal)

Table

Description automatically generated

(Positively Skewed)

1. **Customer segmentation with k-means.** Problem: find a high-quality clustering using K-means and discuss the profile of each found cluster (in terms of the properties that describe the properties of the customers of each cluster). The report should illustrate the adopted clustering methodology and the cluster interpretation. In particular, it is necessary to discuss the identification of the best value of k and the characterization of the obtained clusters by using both analysis of the k centroids and comparison of the statistics of variables within the clusters with that in the whole dataset.

The k-means algorithm is used in rapidminer to cluster the data with the objective of minimizing intra-cluster space. 5 clusters were determined for the k-means algorithm and the algorithm uses the squared Euclidean distance to determine the similarity between the points.

Table

Description automatically generated

The amount k centroids were determined with the k-means++ algorithm which randomly initializes a centroid and takes the furthest data point from the nearest centroid as the next centroid. This algorithm yields a k value of 5. This is a suitable number for k because the cluster is going to be used to identify the consumer and having too many clusters may be result in groups that are too hard to take action on.

Below is the analysis of the individual clusters and their characteristics. Since the attributes are normalized with z-transformation, the characteristics will only tell if this group is below or above the average. Range transformation is not really appropriate because some of the attributes are skewed so using range transformation will still result in a skewed distribution. Only the characteristics that stand out will be pointed out in the analysis below, otherwise the attribute has an average value in the population.

Cluster\_0: Average financial status with low usage of call but high number of over-15-min calls. Considered leaving and over half of the population chose to leave.

(5831 items)

Cluster\_1: Average financial status with high usage of call but low satisfaction. On average, has considered leaving.

(2375 items)

Cluster\_2: Average financial status with lower-than-average usage level. Has high satisfaction level.

(3343 items)

Cluster\_3: Average financial status with high satisfaction and high usage level.

(3857 items)

Cluster\_4: Average financial status with lower-than-average overage charges and low over-15-minute-call. Has low usage level and low satisfaction

(4594 items)

This clustering is validated through the cluster-distance-performance which looks at the average distance between a point and the centroid within a cluster. The overall performance value is -2.817 which is fairly high when the points are on scales between -2 and 2. However, this is to be expected with data of high dimensionality. This performance value could be reduced by increasing the value of k but that would not be too useful in this case because as each cluster becomes smaller, it becomes harder to pinpoint which cluster a new customer would belong to.

The clusters are somewhat reliable because clusters such as cluster 0, with the highest population, can be seen in the data analysis stage as many of the notable attributes overlap in terms of density. Cluster 3 where consumers have high usage level and high satisfaction makes sense as well.

1. **Explain Predictive models (to predict whether the customers churned (STAY or LEAVE))**Illustrate the adopted methodology and the validation and interpretation, describing also the process adopted to select the parameters, together with its quality evaluation.
2. Which predictive method does the best on predicting customer churn?  Provide an explanation that is based on formal evaluation methods. Good samples are k-fold cross-validation of accuracy and ROC curve. You should evaluate your models on dedicated subsets of data that are not involved in training.

Using logistic regression can be reliable at predicting customer churn, especially in this situation where the dependent variable, “churn”, is binary. Upon using logistic regression on the training sample, which is set to be 70% of the given dataset, the logistic regression will yield a correlation between each attribute and the targeted attribute. This can then be used on the test sample to see the reliability of the model. However, using the performance binomial test resulted in an accuracy value of 62.43%.

Table

Description automatically generated

A better predictive method could be to use a decision tree as it works for both numerical data as well as categorical data. It can also eliminate the problem of overfitting by limiting the depth of the tree, which will ultimately result in a more precise prediction model. The minimal gain value is adjusted to 0.02 as to filter out some of the noise and maximal depth is kept at 8 as anything higher would result in overfitting. Depth of 11 starts to result in lower accuracy. Again, the model is evaluated by testing the predictive model, derived from the training set, on the test set.

Table

Description automatically generated

1. Is the data balanced? If not, how does this affect your evaluation?

The data is not balanced. For most of the attributes, the data either follows a bimodal distribution or are positively skewed. This affects logistic regression, as skewed data will skew the regression and result unreliable correlation between attribute and targeted attribute. However, this does not affect decision tree because decision tree only cares about the split in data that brings about the greatest difference between the branches. Thus, the distribution of the data within the attribute does not really matter.

1. What have you done to minimize its impact? (Hint: what is the accuracy of a prediction method that always outputs “No”?)

With the logistic regression method, z-score normalization is used to minimize the impact of the skewed data. Unlike range transformation, which can transform data into a common interval but keeps the skewed distribution, z-score models the distribution of the data based on standard deviation. The average is set as 0 and anything below average is negative, anything above average is positive. This will result in a more balanced distribution, leading to a more reliable regression model.

1. Report the correlations between data entries and customer churn based on your predictive models.

A picture containing diagram

Description automatically generated

Looking at the decision tree above, it is evident that consumer who have house values above 600000 and income below 99800 has a high chance of staying. Consumers who fall into the same category but are being overcharged more than 127 are very likely to leave. In the same vein, consumers who have longer average call durations but have less than 88 leftover are very likely to leave.

Provide recommendations for reducing churn, based on the data.

Based on the correlation above, overcharging customers may be one of the main reason customers are choosing to leave. Hence, proper marketing of other plans with additional call time should be done, so that customers can choose data plans that are right for them. Another correlation is that customers with less than 90 leftover but with over 4.5 minutes average call duration has high chance of leaving. This could be solved by providing plans that cater towards those that prefer longer call duration but less frequent calls so that they don’t have to pay for what they won’t use. This is supported by the k-means clustering algorithm where cluster\_0 shows a similar group of customers (long calls but low overall usage) tend to leave.