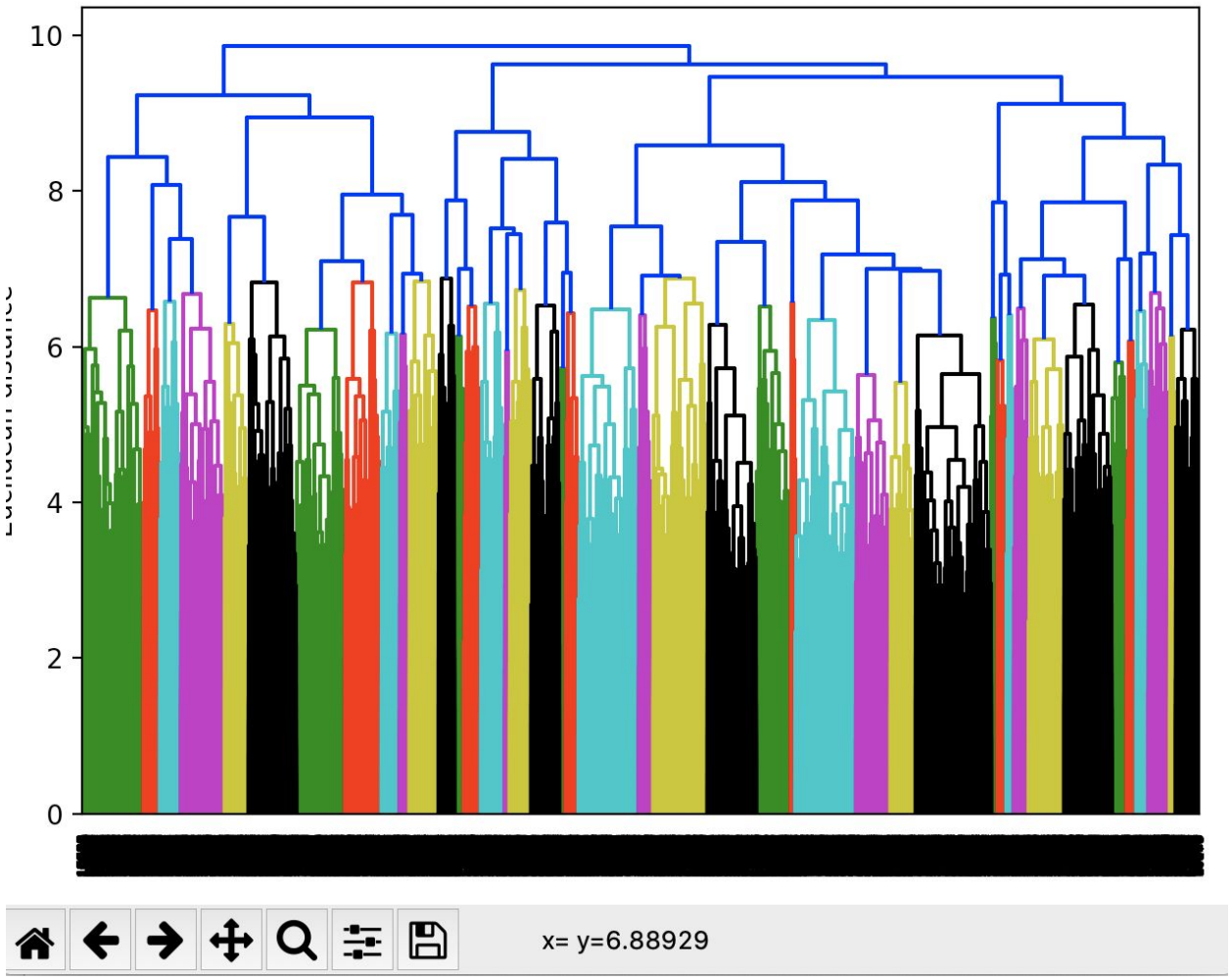
Sreevani Suvarna

Introduction to Data Mining

## Report Analysis:

1. *This data basically takes in 11 attributes: College(whether the customer is in college or not), Income(how much income the customer has), Overage, Leftover, House cost, Handset Price, the number of calls the customer takes that are over 15 minutes, the average call duration, their satisfaction with the company, reported usage level, and whether they were considering changing company. In order to simplify this data, I realized that I had to normalize all the numerical attributes, and simplify the data more by mapping all the ordinal attributes such as reported_usage level to rank numbers and binary attributes(yes/no) such as college to numbers 1 or 0. This is an important step as it would scale the data and normalize it. After normalizing the data, I've created two visualizations for this data. The first visualization is a decision tree which I thought was meaningful because it really lays out the attributes and which combination leads to which output. My second visualization was the dendrogram I produced when I was using K-means clustering to see the similarity between different clusters. I used this dendrogram to measure the distance and get the best k to start with for this algorithm.(I've attached the decision tree in the submission folder and the dendrogram below in the next question).

2. To analyze the clusters using K-means, I coded up the K-Means algorithm using scikit learn in Python. I picked the best k by using hierarchical clustering to first create a dendrogram:




By analyzing this dendrogram, I came to the conclusion that the best k is 6 by cutting off at a distance of 8, and thus used that k value to create 6 different clusters. From producing these clusters, I've come across some surprising results:

**Cluster 1  Similarity**

| | |
|---|---|
| College | zero |
| Average Income | 79009.9 |
| Average Overage | 198.87 |
| Average Leftover | 19.7070 |
| Average House value | 492974 |
| Average Handset Price | 374.0209 |
| Average Overage 15min | 20.8126 |
| satisfaction | Very unsatisfactory |
| usage_level | Same everywhere |
| Considering to change plan | considering |
| Leave or STAY? | LEAVE |

**Cluster 2:**

| | |
|---|---|
| College | one |
| Average Income | 131585 |
| Average Overage | 49.741 |
| Average Leftover | 22.636 |
| Average House value | 487700 |
| Average Handset Price | 687.54 |

| | |
|---|---|
| Average Overage 15min | 3.8930 |
| satisfaction | Very unsatisfactory |
| usage_level | Same everywhere |
| Considering to change plan | considering |
| Leave or STAY? | STAY |

**Cluster 3:**

| | |
|---|---|
| College | zero |
| Average Income | 56657 |
| Average Overage | 36.317 |
| Average Leftover | 7.10303 |
| Average House value | 489632.1 |
| Average Handset Price | 264.241 |
| Average Overage 15min | 2.78168 |
| satisfaction | Very unsatisfactory |
| usage_level | Same everywhere |
| Considering to change plan | considering |
| Leave or STAY? | STAY |

**Cluster 4:**

| | |
|---|---|
| College | one |
| Average Income | 67737.552 |
| Average Overage | 49.6397 |
| Average Leftover | 16.0705 |
| Average House value | 494608.44 |
| Average Handset Price | 314.9060 |
| Average Overage 15min | 3.893617 |
| satisfaction | Very unsatisfactory |
| usage_level | Same everywhere |
| Considering to change plan | no |
| Leave or STAY? | STAY |

**Cluster 5:**

| | |
|---|---|
| College | zero |
| Average Income | 63096.68 |
| Average Overage | 56.0871 |
| Average Leftover | 62.598 |
| Average House value | 499920.30 |
| Average Handset Price | 288.36 |
| Average Overage 15min | 4.3647 |
| satisfaction | Very unsatisfactory |
| usage_level | Same everywhere |
| Considering to change plan | considering |
| Leave or STAY? | LEAVE |

**Cluster 6:**

| College | one |
|---|---|
| Average Income | 80321.3 |
| Average Overage | 84.7873 |
| Average Leftover | 23.336 |
| Average House value | 496817.74 |
| Average Handset Price | 390.442 |
| Average Overage 15min | 7.5107 |
| satisfaction | average |
| usage_level | Same everywhere |

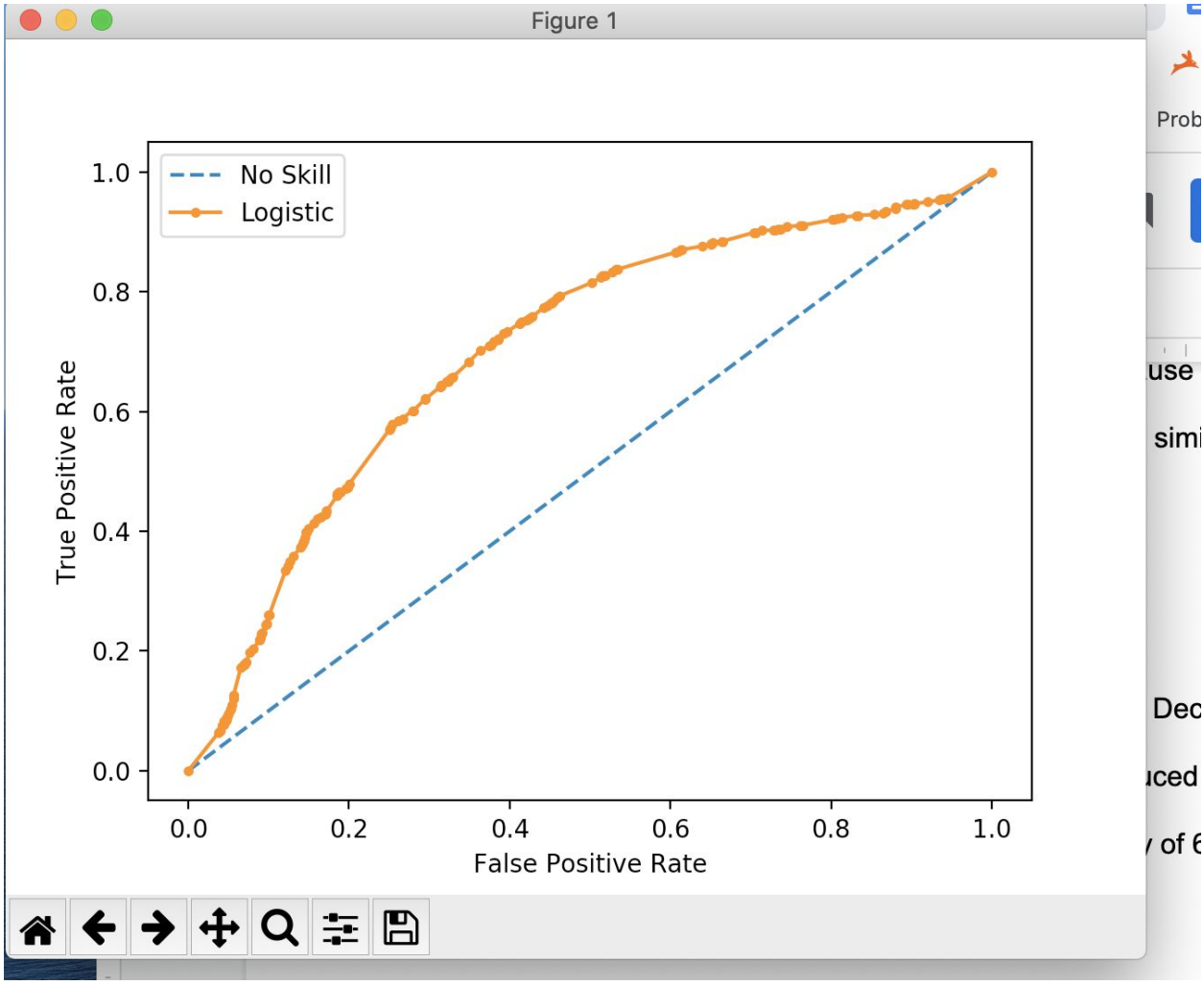
| | |
|---|---|
| Considering to change plan | considering |
| Leave or STAY? | STAY |

**What differences did I see across the clusters?**

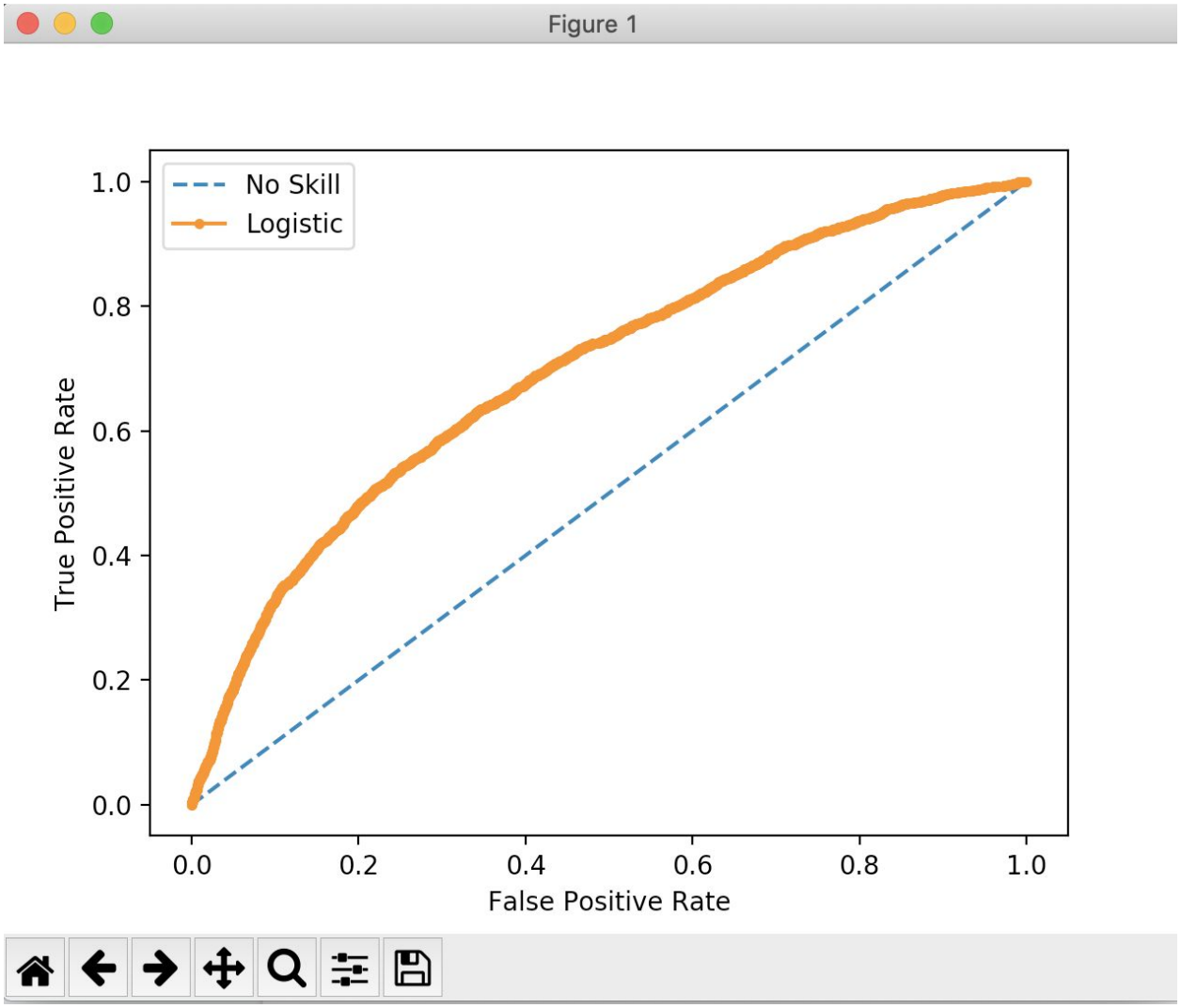
I saw that roughly all the customers in all the clusters had almost the same average of calls. However, the income differed majorly across all the clusters: for example, in cluster 1 the average income was $131545.6 whereas the average income in cluster 2 was $56946.2. Another major difference across the clusters was whether the customers churned to stay or leave. Some clusters were similar because most of the customers wanted to leave, whereas some of the other clusters were similar because most of the customers wanted to stay.

3.

a. I coded up two predictive models for this dataset: a Decision Tree model and a Logistic Regression model. The decision tree produced an accuracy of 68% and the logistic regression model produced an accuracy of 65%. When I produced the ROC curve for the Logistic Regression model and Decision Tree it looked somewhat like this:




**This is the ROC curve for Decision Tree**




**This is the ROC curve for Logistic Regression**

As I can see, the ROC curve is not touching the top left corner : it's almost there but not yet. However, I believe that since the Decision Tree produced slightly a higher accuracy, I would say that the Decision Tree method works better on customer churn.

b. I don't think my data is balanced because when I coded up the KMeans : the clusters varied highly and in addition, the mean was always higher than the median which means that the data might have been skewed a little. In addition,

when I saw the Decision Tree , I saw that the model was very overfitted which means that the data wasn't balanced.

c. To balance the data, I wrote a method in python that maps all the categorical variables to a number. To elaborate, I mapped the attributes: College, Reported Unsatisfactory, the labels(Leave or Stay), Reported Satisfaction, Considering Change of Plan all to a value. For example, I mapped the values in College to 0 or 1 : 0 for zero and 1 for one and the values in Reported Satisfactory to values 0(very unsatisfactory), 1(unsatsfactory), 2(satisfactory), and 3(very satisfactory).
d. Using my decision tree model, I've come across that lower house rates lead to customers staying more, whereas higher income and house rates correlate with customers leaving the company. Also with my logistic regression model I printed

out the correlations as shown below:

| | |
|---|---|
| COLLEGE | 0.118135 |
| INCOME | -0.163843 |
| OVERAGE | -0.411277 |
| LEFTOVER | -0.246192 |
| HOUSE | 0.461332 |
| HANDSET_PRICE | -0.087725 |
| OVER_15MINS_CALLS_PER_MONTH | -0.134970 |
| AVERAGE_CALL_DURATION | -0.124559 |
| REPORTED_SATISFACTION | -0.010970 |
| REPORTED_USAGE_LEVEL | -0.020736 |
| CONSIDERING_CHANGE_OF_PLAN | -0.037383 |

As we can see income, overage, leftover, handset price, over 15 minute calls, average call duration, reported satisfaction, reported usage level, and considering change of plan all have a negative correlation. Whereas, college, and house value have a positive correlation with customer churn.

e. There are many recommendations for reducing churn. By looking at my clusters, it seems that many customers were very unsatisfactory with the company, and most of them didn't have that much of an income. To reduce churn, the company could provide more grants or more deals to get more opportunities within the

company. In other words, the company could give out more discounts with their products or monthly plans to the customers. Another factor could be to send out an outline on how each customer's bill would look like before they are actually charged to let them know what their costs consist of. Looking at my decision tree, it looked like customers who mostly had over 15 minute calls and more usage of the phone had a higher chance of leaving. Hopefully, sending out an earlier outline of their plan would let them change their behavior towards the phone so their bills don't be costly.