CSE 145 - Homework 2

Daniel Xiong (dxiong5@ucsc.edu)

Due May 14

1 Introduction

The goal of this project was to data mine the `Customer_Churn.xlsx` dataset, which contains 20,000 entries with 12 variables describing features of customers of a mobile phone provider. We aimed to predict the variable `"LEAVE"`, which represented whether a given customer would stay or leave the company. To achieve this goal, we first create some visualizations to try and understand the data. We then used the k-means algorithm to cluster the data to further analyze the properties of the data. Finally, we chose predictive models to predict whether or not a customer would stay or leave the company.

2 Tools Used

This assignment was completed using Python 3.7 (`scikit-learn` >= 0.22.1, `pandas`, `matplotlib`).

3 Data Understanding

One important metric is information gain, which is a measure of how much an attribute improves entropy over the whole segmentation it creates. In the context of supervised segmentation, information gain measures the knowledge gained by splitting the set on all values of a single attribute. **Figure 1** is a bar graph that ranks the information gain of all the attributes in decreasing order, with `"HOUSE"` `"INCOME"` being the attributes with the greatest information gain values. The other attributes had significantly lower information gain.

write about next visualization

The code for these visualizations can be found in `visualizations.py`.

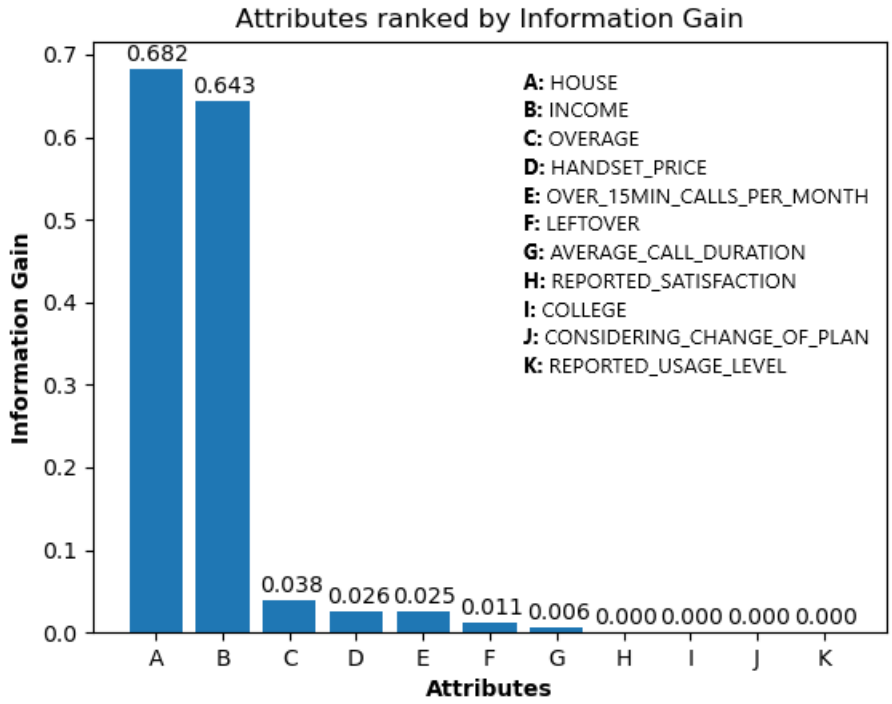

Figure 1: Attributes ranked by their information gain

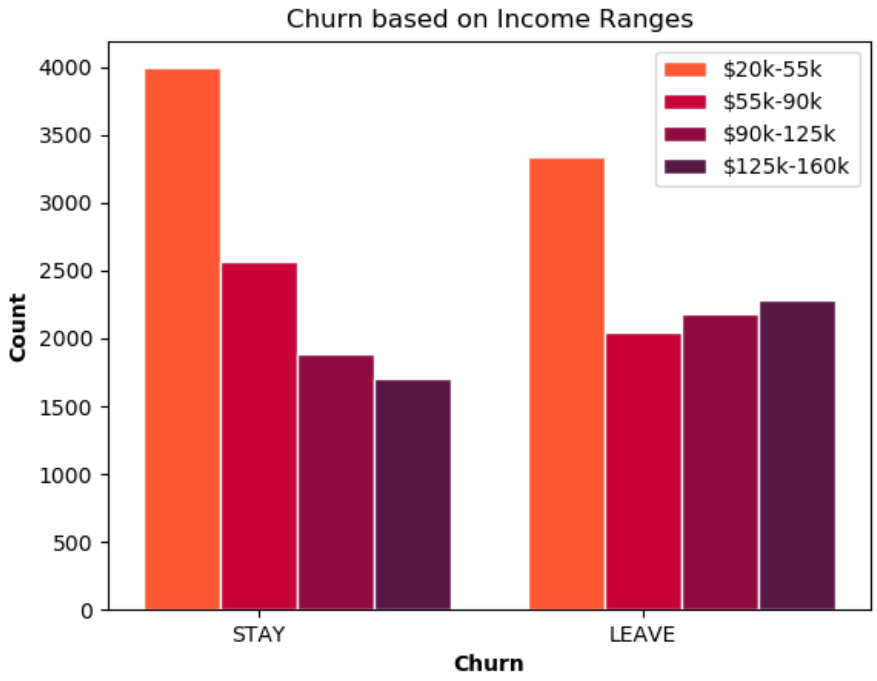

Figure 2: Income vs Churn

4 Customer Segmentation with k-means

Before training a k-means model on the dataset, I first had to scale the data so that the magnitudes would not be so vastly different. To do this, I used the `StandardScaler` function from Python's `sklearn` module. I then used `sklearn`'s KMeans function for the k-means model along with the `k-means++` centroid

initializer.

In order to determine the best value of k, I trained many different k-means models each with a different k from $k = 2 \rightarrow 25$. I then created two plots: a Sum Squared Error (SSE) plot and a Silhouette plot, **Figure 3** and **Figure 4**, respectively. I used the elbow method, along with the silhouette plot, to determine that a k-means model with $k = 5$ would result in the optimal clustering.

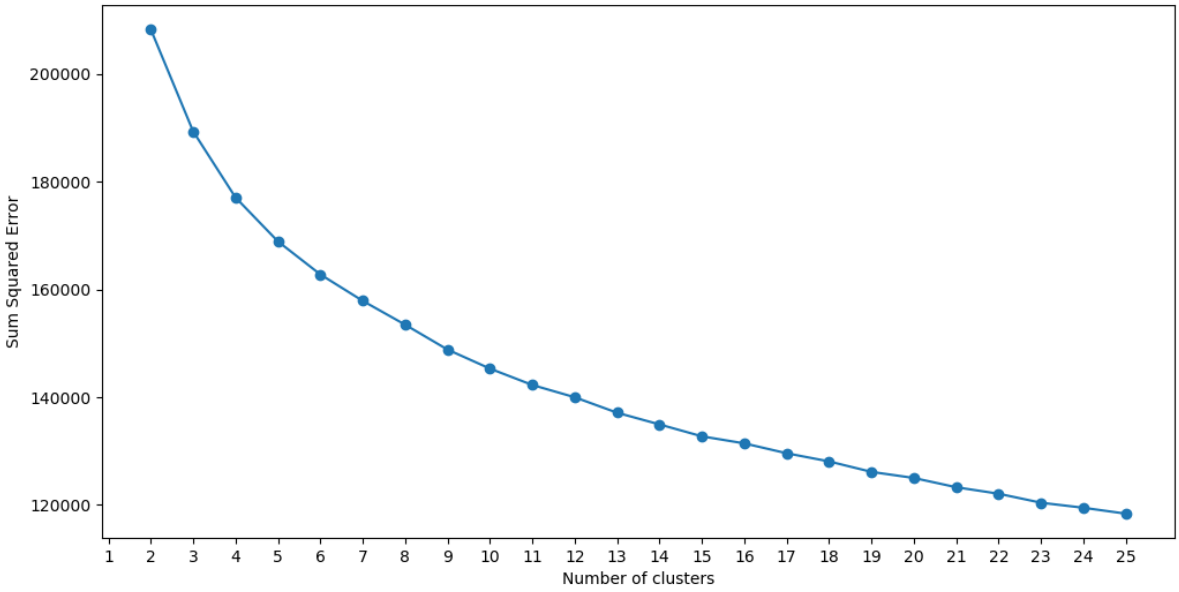

Figure 3: SSE plot

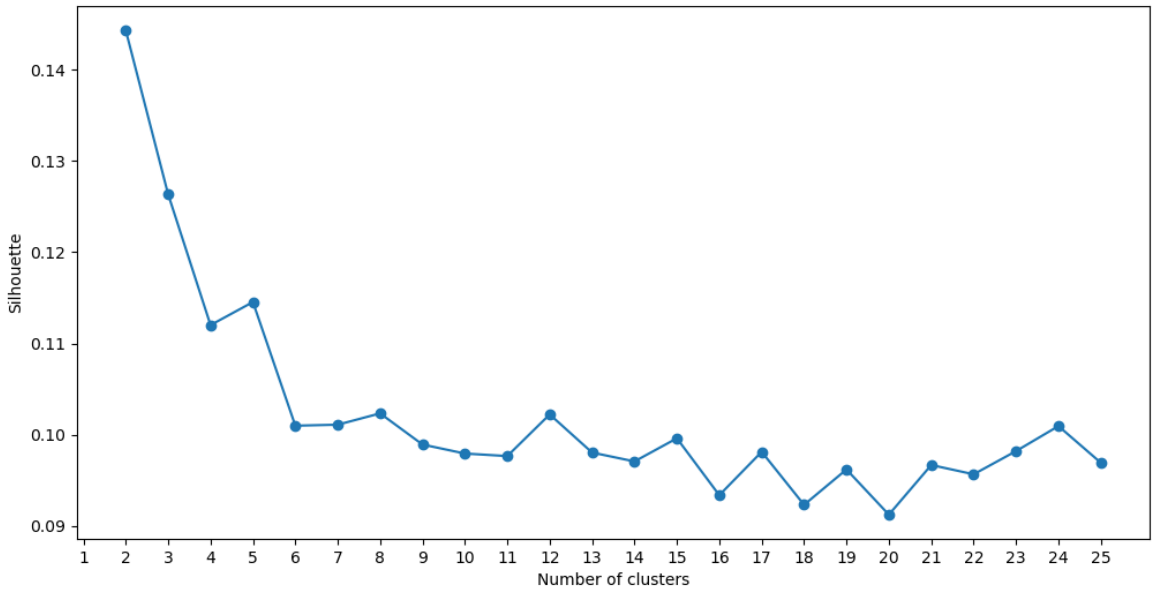

Figure 4: Silhouette plot