Darren Yu

TIM 147 - HW4

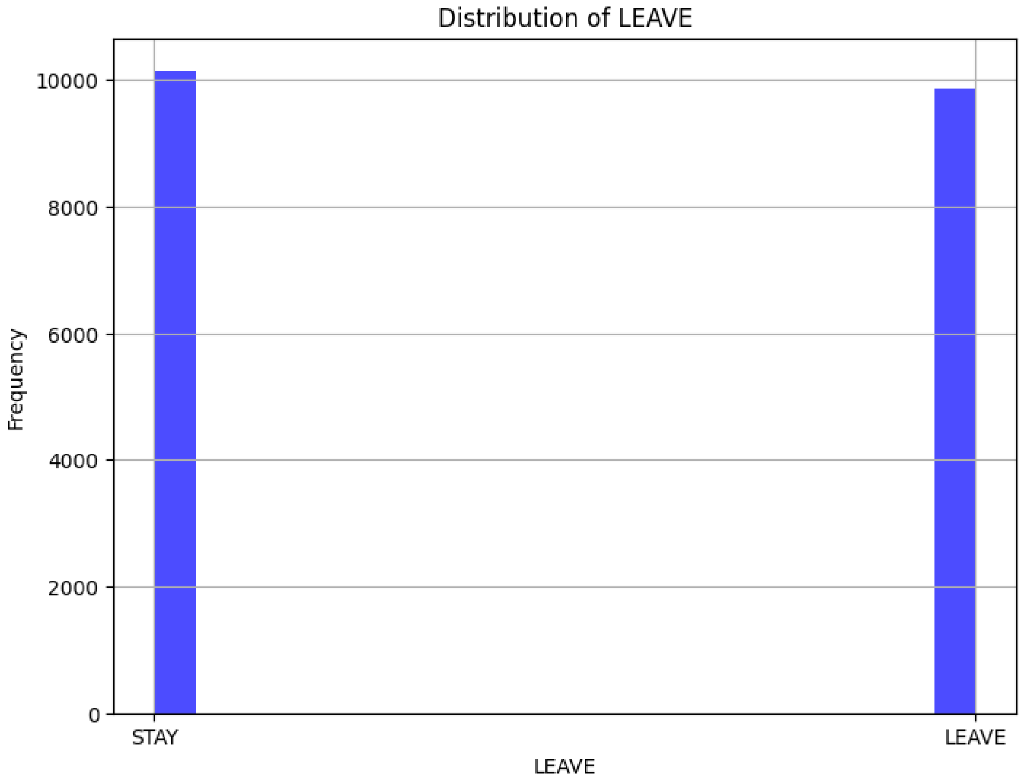
Data Understanding (DATA SET REMAINED THE SAME FROM LAST LAB) :

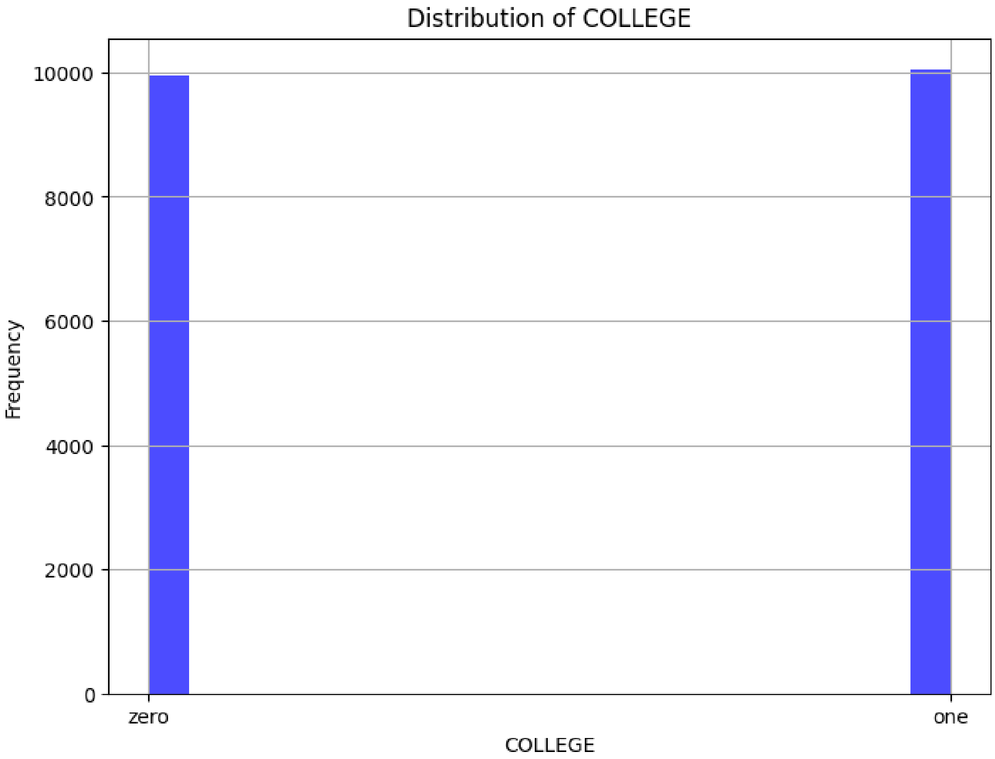
- The Data consists of 12 columns each an attribute of a Customer who partakes in a company, the Columns are as follow: COLLEGE, INCOME,OVERAGE,LEFTOVER HOUSE,HANDSET_PRICE, OVER_15MINS_CALLS_PER_MONTH, AVERAGE_CALL_DURATION, REPORTED_SATISFACTION, REPORTED_USAGE_LEVEL,CONSIDERING_CHANGE_OF_PLAN, LEAVE
- **Distribution Analysis for category LEAVE**: There are about 10000 consistent decisions made by certain customers on whether or not they chose to LEAVE, or STAY within the company. The frequency of LEAVEs had more counts than the STAYs.
  Two visualizations that are helpful to show this frequency are the histograms and box plots. As for the statistical analysis of the category only so much can be said because they are qualitative values that are ordinal numerical types of data, meaning that they only have the option of being “STAY” or “LEAVE” so we can say that the frequency of the LEAVES is greater than the frequency of the STAYS based on the histogram totals.
- **Neural Networks,** we perform a customer churn between the binary target of leave and compare it with the other classifications we had available in the data set. This is done by using the chi-squared tests to determine whether there is a significant correlation between each predictor and customer churn. We then iterate through categorical columns, creating contingency tables and calculating chi-squared statistics and p-values. It then prints the results, indicating whether each predictor has a statistically significant correlation with

customer churn (p-value < 0.05) or not. This analysis helps identify which categorical variables are associated with customer churn, providing insights for improving customer retention strategies.

- Management of Missing Variables: We ran a missing variable search on the data to check for any gaps within the inputs, from what we can see the data is fully available with no missing values.

2 visualizations:

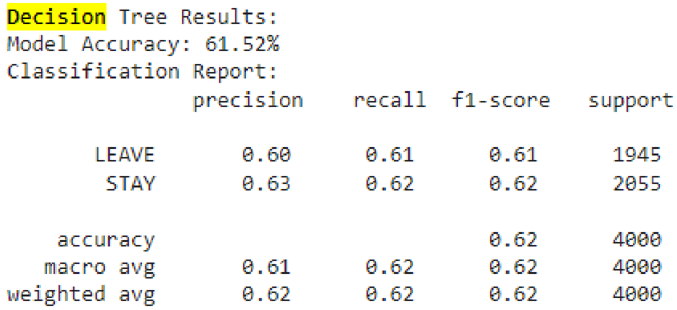




Both of these histograms show similar shapes in terms of their distributions. Might be able to determine a correlation between going to college and leaving the company for some reason.

Predictive Methods and Findings:

1) Decision tree

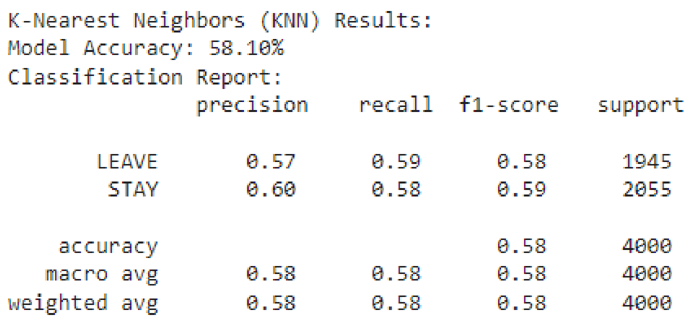


```
Decision Tree Results:
Model Accuracy: 61.52%
Classification Report:
              precision    recall  f1-score   support

       LEAVE       0.60      0.61      0.61      1945
        STAY       0.63      0.62      0.62      2055

    accuracy                           0.62      4000
   macro avg       0.61      0.62      0.62      4000
weighted avg       0.62      0.62      0.62      4000
```

2) Knn predictive model



```
K-Nearest Neighbors (KNN) Results:
Model Accuracy: 58.10%
Classification Report:
              precision    recall  f1-score   support

       LEAVE       0.57      0.59      0.58      1945
        STAY       0.60      0.58      0.59      2055

    accuracy                           0.58      4000
   macro avg       0.58      0.58      0.58      4000
weighted avg       0.58      0.58      0.58      4000
```

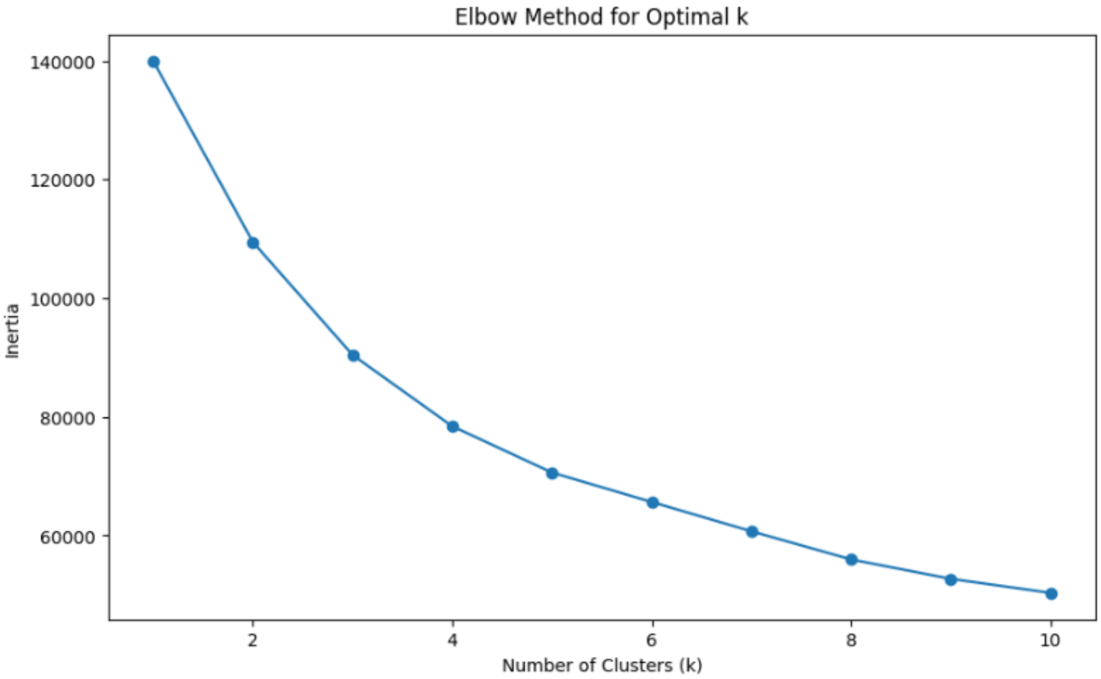
Which predictive method does the best on predicting customer churn?

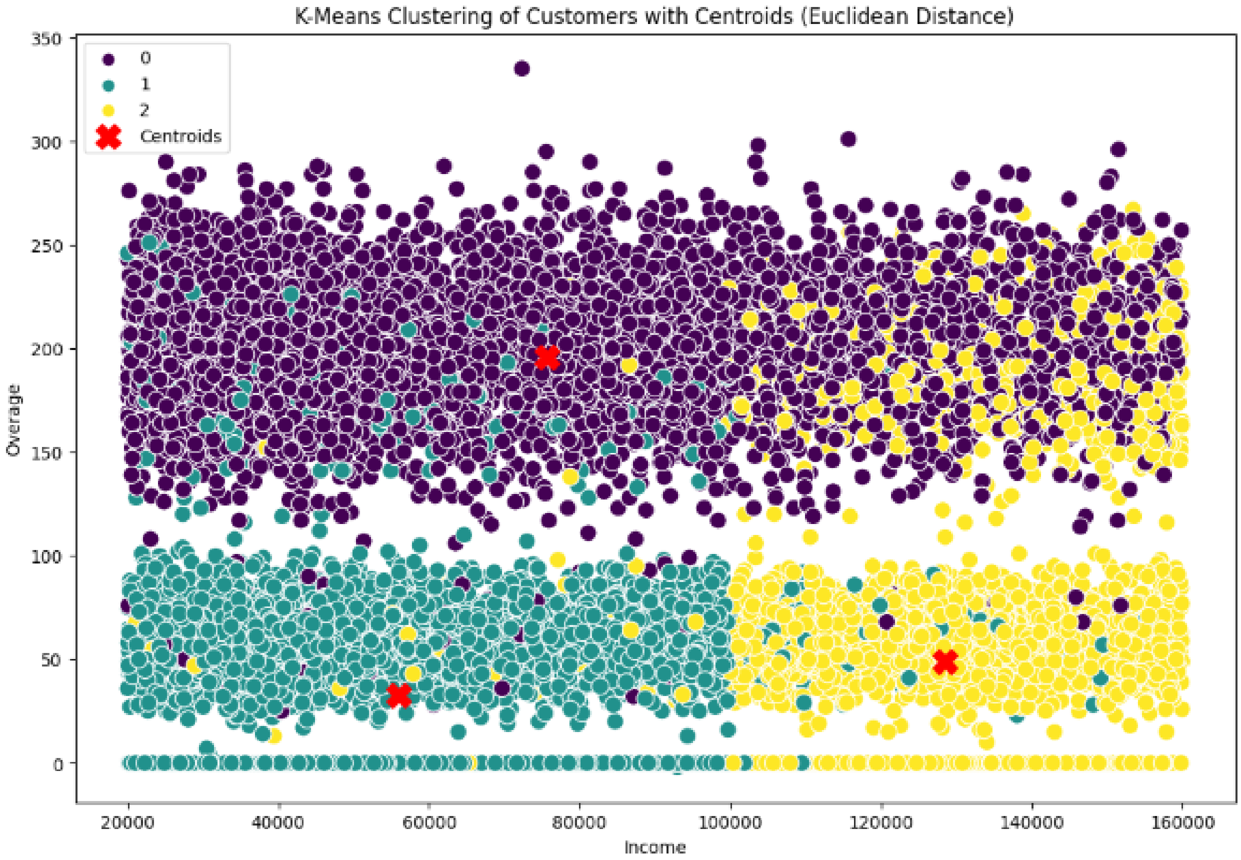
The accuracy of the Decision Tree was more on target than the accuracy of the KNN model.

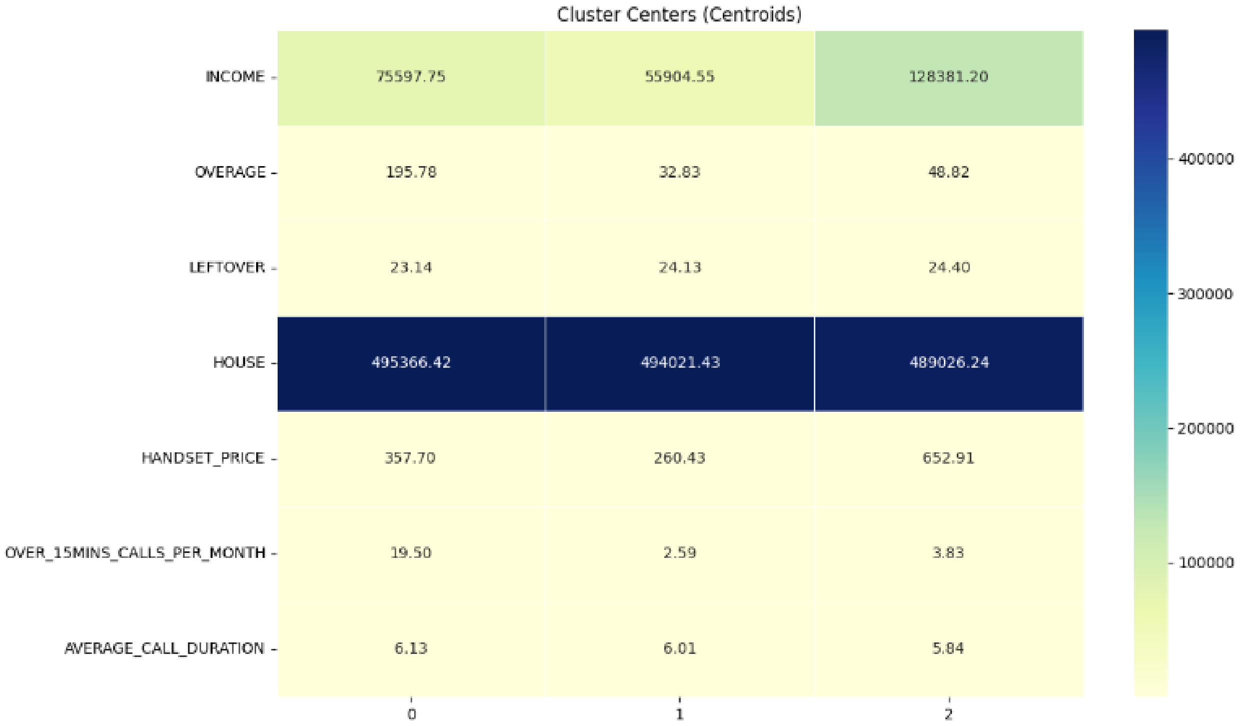
- Model Accuracy for the Decision Tree was 61.25%
- Model accuracy for the KNN model was 58.10%
- The precision number for the Decision Tree was scored at:
  - .60 for LEAVE variables
  - .63 for STAY variables
- The precision numbers for the KNN model was scored at:
  - .57 for LEAVE variables
  - .60 for STAY variables.

These results provide insight into the model accuracy and better churn coming from the Decision Tree model over the KNN model, as a grouping based method would sway our churn slightly by .04 on the precision scale as shown above. Given the accuracy is only 3% off from each other, the models, we can say that in terms of the two models the Decision Tree is more accurate although our results are not confident in the 60% accuracy model.

Customer segmentation with K-means :







K-mean analysis:

In order to create the k-mean plot, I utilized the elbow method. This visually informed me that the optimal amount of groupings is three. One drawback is that a lot of the data that is given to us was in words. As in, a large proportion of out data consisted of non-numerical data.

To find the optimal k-value, I simply used the elbow method and initialized inertia as an array. From there, I used a for loop to help me iterate multiple times to find a good centroid; I set a cap at around 300 iterations and plotted the figure. This ultimately gave me a value of k = 3. However, I still hold lots of reservations based on the amount of our data set that isn't numerical. I tried to find and replace stay and leave, but it isn't feasible to create meaningful k-mean plots with binary data, as it is unable to be plotted. In addition, customer satisfaction was given in verbal opinions. I suggest that those categories should be converted to integers or some type of numerical scale, as it is very hard to correlate it without our customer churn and possible business decisions. However, I deleted it because I wasn't sure if I could modify our data set.

Which predictive method does the best on predicting customer churn? Provide an explanation that is based on formal evaluation methods.

I believe that the neural network from the last lab and the decision tree diagrams were much more revealing in terms of what a next move could be. Below, I check with roc curve and above I have metrics to measure the accuracy of my decision trees and my knn model that can help me validate whether the models are sufficient or not. I may not have been able to get the margins to reach a standard of acceptability, but I am able to see what tolerances are. In this case with the k-mean, I don’t like how everything is visual. But in general, it uses the data as the other models we’ve deployed. Because our data is largely non-numerical, it is really hard to make an informed decision because our goal is to maintain and grow the customer base.

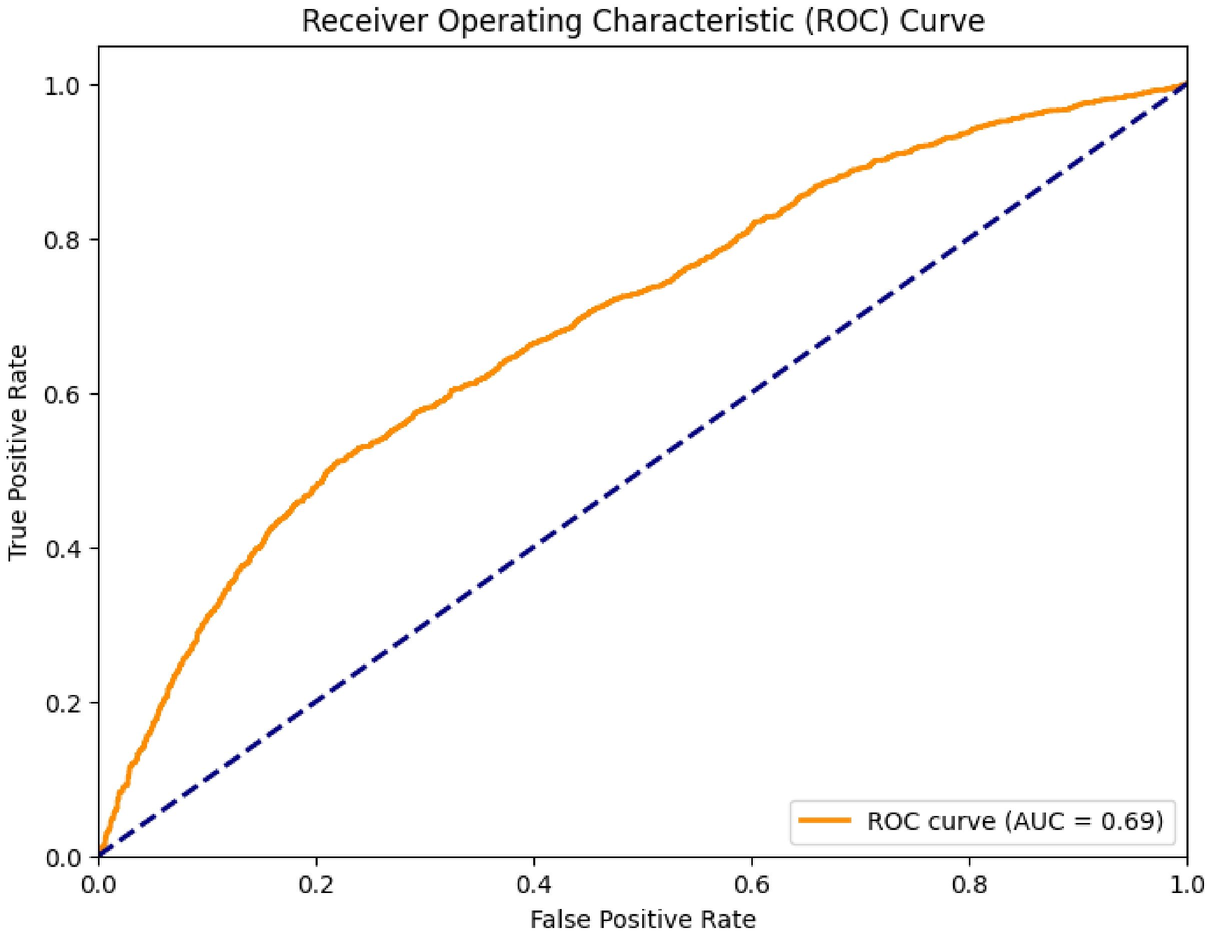
Overall my criticism is that our means of validation is running the model until the results don’t change and there aren’t many precise ways to validate the metrics like the others.

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

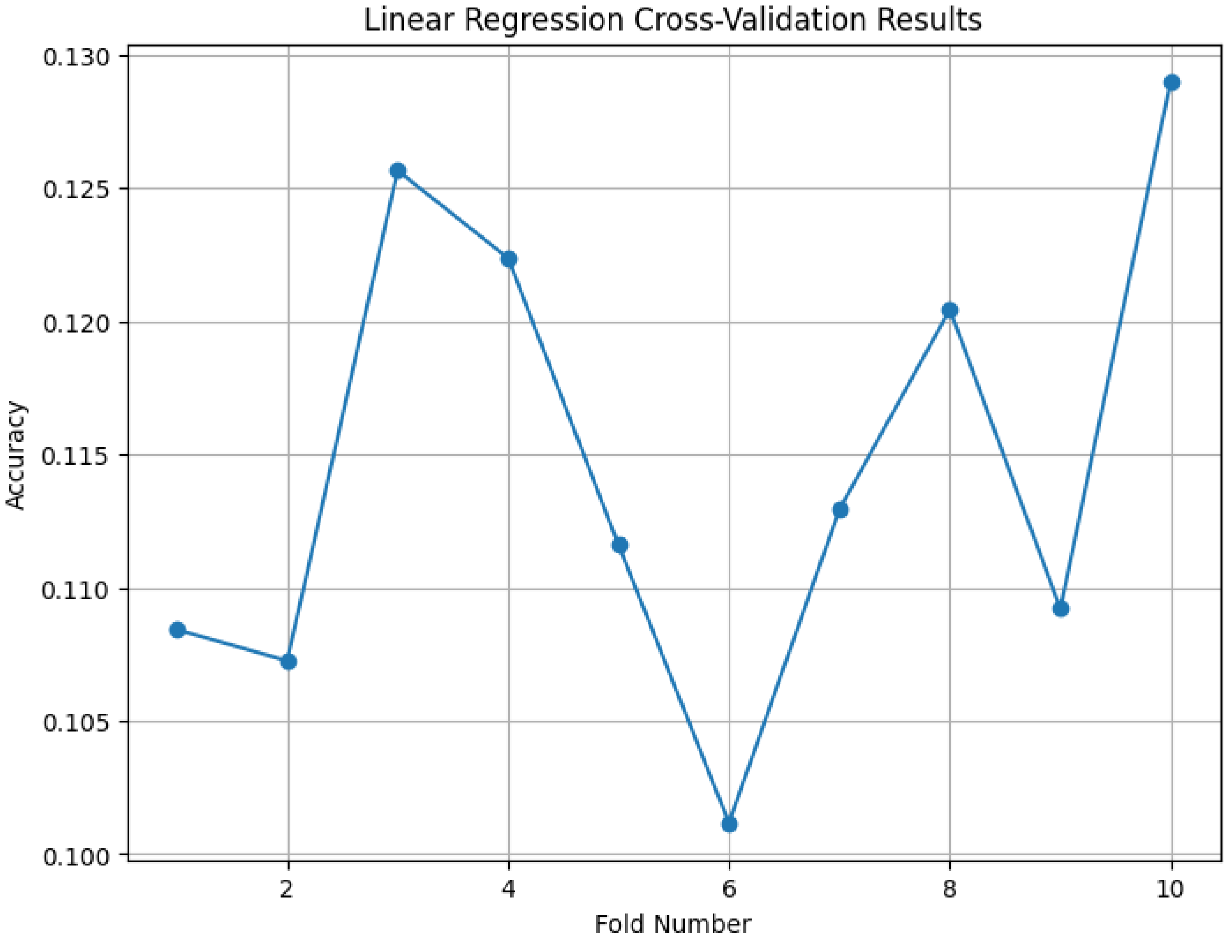
From the histogram and the model findings, we can see that both results were roughly about the same. As in, within 3% of each other. And with further inspection, the precision numbers of the knn and decision tree models are within 0.03 of each other. From the lack of variance, the tendencies and nature of both models do not seem heavily skewed toward leaving or staying.

Validation

ROC curve:



K-Fold Cross-Validation



Insight:

Since the result from both models was extremely similar, I used the linear regression model as the basis for inspection. From our ROC curve, we can see a value of about .69, which is fair but not ideal. But from this, we can see that the data is not discriminatory. As for the K-fold cross validation, we are still confused because the accuracies are such low values; despite, the values shown in our models.

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

Based on the two different models, for the Decision model the accuracy for the LEAVE variable varied around 60% to 61%. For the KNN model, the accuracy for the LEAVE variable varied around 57% to 58%. In order to minimize the impact we could have used less factors in determining the customer churn such as removing variables like if the customer had a house or not or if they went to college. However, that would make the model less accurate in terms of predicting the customer churn as it would force the model to use a smaller version of the data set and would only provide an outcome based on the smaller data set. The larger the data set would provide the more accurate and concrete we can get with determining the different factors that play into the customer churn accuracy.

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

We ran a correlation test between the other data variables in the Customer churn analysis, rating the correlation using the CHI squared statistic which quantifies the difference between the observed frequencies of events and the expected frequencies, assuming that there is no association between the variables. We also used a p value to rate the correlations, the p value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis is true. In other words, it quantifies the strength of evidence against the null hypothesis, being the test variable of LEAVE.

Based on the results, our k-value is three significant groups. Mainly, Income, house, and handset price. However, I failed to include non numerical data in this grouping. From our results, I think they are much more telling of the situation, as our predictive model models are saying that most of our categories are deemed significant despite having drastically different values.

Results:

Correlation Analysis:

COLLEGE:

Chi-Squared: 4.19

P-Value: 0.0407

Statistically significant correlation.

INCOME:

Chi-Squared: 18564.02

P-Value: 0.4490

No statistically significant correlation.

OVERAGE:

Chi-Squared: 1468.37

P-Value: 0.0000

Statistically significant correlation.

LEFTOVER:

Chi-Squared: 449.37

P-Value: 0.0000

Statistically significant correlation.

HOUSE:

Chi-Squared: 19685.93

P-Value: 0.5309

No statistically significant correlation.

HANDSET_PRICE:

Chi-Squared: 990.08

P-Value: 0.0000

Statistically significant correlation.

OVER_15MINS_CALLS_PER_MONTH:

Chi-Squared: 987.69

P-Value: 0.0000

Statistically significant correlation.

AVERAGE_CALL_DURATION:

Chi-Squared: 250.00

P-Value: 0.0000

Statistically significant correlation.

REPORTED_SATISFACTION:

Chi-Squared: 8.56

P-Value: 0.0730

No statistically significant correlation.

REPORTED_USAGE_LEVEL:

Chi-Squared: 1.02

P-Value: 0.9062

No statistically significant correlation.

CONSIDERING_CHANGE_OF_PLAN:

Chi-Squared: 3.36

P-Value: 0.5001

No statistically significant correlation.

LEAVE:

Chi-Squared: 19996.00

P-Value: 0.0000

Statistically significant correlation.

Provide recommendations for future action:

- Prioritize efforts on factors that have shown statistically significant correlations with customer churn, such as college, coverage, leftover, hand set price, over fifteen minute calls, and better scoping for average call duration. Since these are the most likely to influence customer churn, it would beneficial to the company to address the concerns of this particular customer segment first
    - For coverage, it may be beneficial to look at where the highest concentrations of bad coverage are and deploy infrastructure to resolve the issue
    - The business could also restructure their pricing to ensure they align with the customer's expectations. If you pay less, people might not mind the current services.
- Competitive analysis may also be a good exploration for the company.
    - Since the company is concerned about losing customers, try to learn what other companies are doing to make them want to jump ship
    - It would be wise to focus on what the customers group's we've indicated as likely to switch to a competitor

- Provide recommendations for future action:
- Prioritize efforts on factors that have shown statistically significant correlations with customer churn, such as college, coverage, leftover, hand set price, over fifteen minute calls, and better scoping for average call duration. Since these are the most likely to

influence customer churn, it would beneficial to the company to address the concerns of this particular customer segment first

- For coverage, it may be beneficial to look at where the highest concentrations of bad coverage are and deploy infrastructure to resolve the issue
- The business could also restructure their pricing to ensure they align with the customer's expectations. If you pay less, people might not mind the current services.
- Competitive analysis may also be a good exploration for the company.
- Since the company is concerned about losing customers, try to learn what other companies are doing to make them want to jump ship

Group Repository

https://github.com/YuMe-02/TIM_147.git

GPT searches

https://chat.openai.com/share/06232861-0bd2-4d58-a00d-97ca612c0854

https://chat.openai.com/share/d5620173-d877-49aa-abcf-e3402cd06534