Fyez Dean

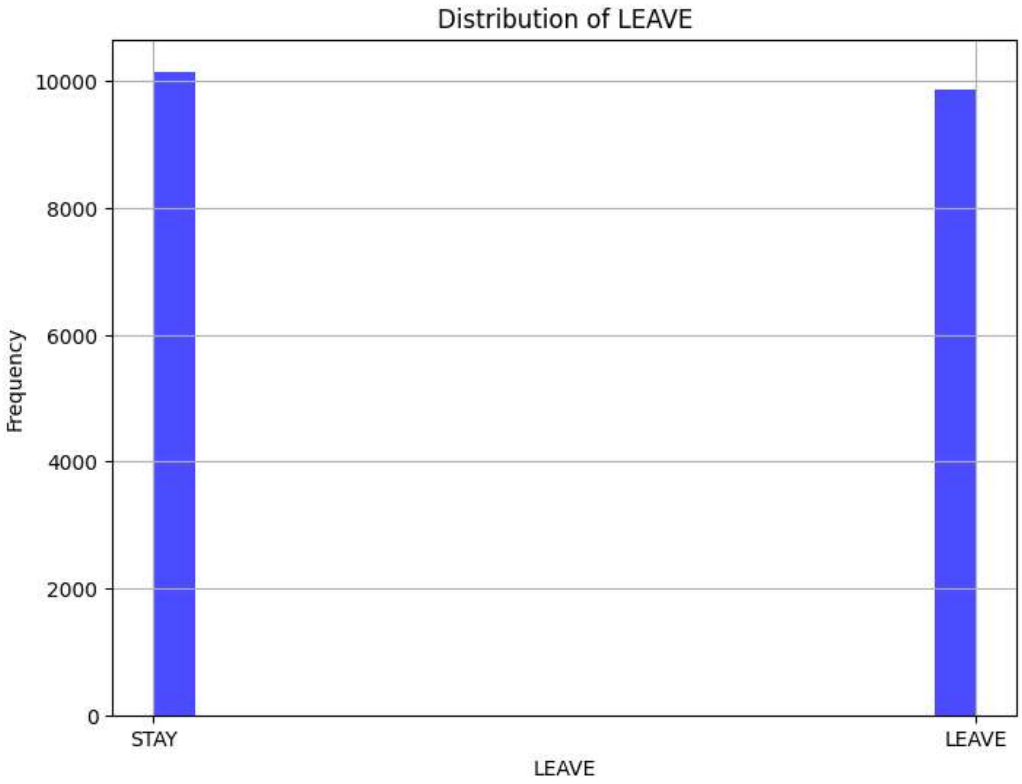
Darren Yu

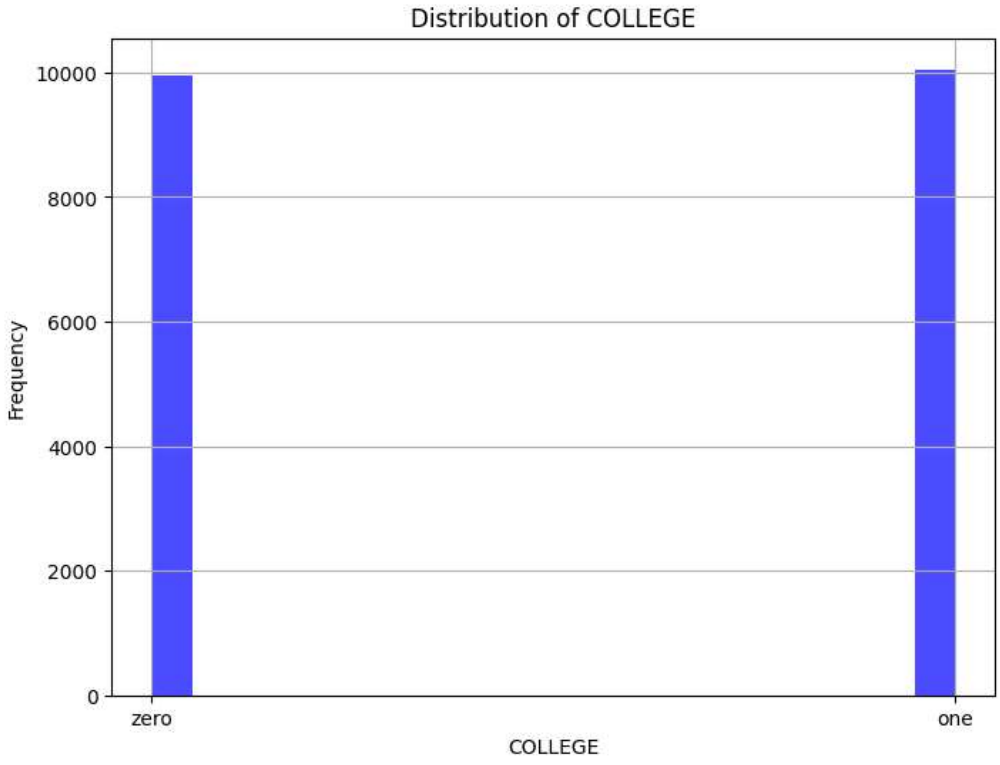
TIM 147 - HW2

Data Understanding:

- 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.
- 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.

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.

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.

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 reducing churn, based on the data.

1) **Pruning:**
    - Pruning can prove beneficial to the predictive modeling process because it can reduce and get rid of unnecessary groupings that might over complicate the model. Thus, resulting in simpler graphical representation and reduced propagation of errors that occur with a larger scale of groups.
2) **Gini:**
    - This rating can be used to help classify groups and assist in the pruning process. By checking for the purity of certain data points, we can reduce noise and outliers. By carefully observing points and sifting, we can generate more accurate predictions. Overall, it provides clarity in the closeness and relationship between subgroups.

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