Exploratory Data Analysis Project

Jacob Polomsky

June 8, 2024

A1 . Does Income affect customer churns?

A2. This analysis will help to see if income is a determining factor on whether customers Churn or not.  If this is the case it might benefit the company by looking at customer that have high churn over rate based off income and either finding way to incentivize them to stay or when doing marketing finding ways to appealing to customer that are more likely to stay.

A3. The data that we are going to be using for the T-test is Income and Churn.

B1- See ipynb File attached

B2- T- Test

A screenshot of a computer

Description automatically generated

A white box with black and green text

Description automatically generated

B3- The reason why I did the a independent T-test is because we are using only one categorical variable and one continuous variable. We don’t use chi squared because that requires two categorical variables. For Anova and T-test they both can use one categorical variable and one continuous, but because we are using Boolean like values we go with T-test because Anova recommend to use three or more categories and while the T-test uses only two.

C. The values I decided to choose for my univariate statistics for my Categorical I choose Income and Area, and for Continuous was Age and Income. Age has close to a constant distribution with almost all values falling within 160 to 140.

A number on a white background

Description automatically generated

Income has a left skewed distribution with a lot of the values falling on the lower end and less with the higher values and the mean being about 40,000

A screenshot of a computer

Description automatically generated

Area has a even count of variables all areas being between 3300 and 3350

A screenshot of a computer

Description automatically generated

Churn has and uneven distribution with having a lot more customers that have not churn than that have churned

A screenshot of a computer

Description automatically generated

C1. see word doc attached for visualization

D. For the Bivariate Statistics I group Age and Income, Churn and Area. With Income and Age I took the mean for each age and found that the relationship is close to constant.

A table with many small colored numbers

Description automatically generated with medium confidence

For Churn and Area, I found the these also were close to constant with No being between 2500 and 2400 and Yes being 925 and 860.

A screenshot of a computer

Description automatically generated

D1. see word doc attached for visualization

E1. So, for our hypothesis we are testing whether income has an effect on whether customers churn or not. Our null hypothesis is that income does not have an effect on customer churn. We set our significance level to 0.05, below the value we reject the null hypothesis and above that value we accept the null hypothesis. The p-value we got back from the t-test is .55 which is way above the significance level so we fail to reject the null hypothesis or we accept the null hypothesis as there is no correlation between income and churn.

E.2, One limitation in the analysis that we have done is that we did not factor in income to current charge. It might be true that income does not play a big part, but maybe income to how much the client is charged might have a bigger effect on whether a customer churns or not. It also over looks at competing companies there might be another company that has cheaper cost so that is why customer are leaving or maybe our data speeds are not the fastest and that might be another reason why the customer are leaving.

E.3 As for our result from this test, I would recommend that we take a different direction as to why customer are leaving than Income. I would recommend mix different factors together to see if customer churn like mixing Age and Area, or mixing income to cost have a greater effect on customer churn. I would not recommend the company use income as the main factor in determining whether or not a customer stays or not.