Logistic Regression Project

Jacob Polomsky

June 19, 2024

A1. What factors influence Churn ?

A2. My goal into this analysis is to is to gain better understanding of what variables correlate to Churn.

B1. Four assumptions to logistic regression, First is that the variables are Binary**.** Second making sure there is no multi-collinearity or that one independent variable does affect another independent variable. Third is a sufficient sample Size. Four that each observation is independent of each other.

B2. The two benefits of Python for me is one I am more comfortable with the languages than with R. Python is a general purpose language making it more versatile than R allowing you to do a lot more with it. Python also has a extensive list of libraries that can help with running linear regression. The packages I ran in this analysis is Pandas for easy use of for the dataset. Matplotlib for making visualizations. Statmodels for running the logistic regression model and seeing the summary statistics. Sci-kit learn for the confusion matrix and accuracy\_scores.

B3. The reason why I am running a multiple Logistic Regression is because the Churn is a Categorical variable which makes Logistic regression better predictor than linear regression. Especially since the value is binary a linear model will go outside the bounds of 0 or 1 insinuating that a outcome will happen when that is not the case and mathematically not possible either like it not possible that there is a 150% chance a person churns.

C1. For the Data cleaning process the first thing I did was to narrow down the dataset to use only the columns I needed for the analysis. Then I checked for duplicates to makes sure each observation is independent and found that there were no duplicates in the data set.

A screen shot of a computer

Description automatically generated

The second thing I did was check for null values because the model cant take null values and found over 2000 null values for Internet Service and drop those values.

A screenshot of a computer program

Description automatically generated

A screen shot of a computer screen

Description automatically generated

I used pandas .describe() function to check for outliers to makes sure the data model runs smooth and did not find any evident outliers in the data.

A screenshot of a computer

Description automatically generated

C2 . See other word Document or Logistic\_Regression\_Data\_Exploring.ipynb it will have a more in-depth screenshot \ summary for all categorical variable

A screenshot of a computer screen

Description automatically generated

A screenshot of a black screen

Description automatically generated

C3. See code attached or Logistic\_Regression\_Data\_Exploring.ipynb or other word Doument( I recommend the .ipynb files it will look more organized)

C4. For Data Wrangling I had two sets of categorical variables one having two categories and another having more than two categories. If it had a yes no question, I converted yes to 1 and no to 0 with the replace function. Categorical variable that were not yes or no I used pandas .get\_dummies function to convert those values to into dummy variables.

C5 See csv attached

D1

A screenshot of a computer screen

Description automatically generated

D2

After running both models reduction technique, they both came out to a similar result with accuracy being around .90, precision being around .82, recall being around .79 and f1 being around .80. So, there is no real benefit in either analysis that makes any real difference for this analysis. I picked the method that was easier to read and write and was less computationally extensive Backward Feature Selection. It might not be as good in the long run as a model evaluation metric or it might be better because it might not run into collinearity issues or arbitrary criteria, but for the sake of this model this no benefit or detriment for this analysis for either model. Backward Feature selection runs a model and looks at the p-values of all independent variables and how they affect the dependent variables if the p value is low it shows that this value is significant for the analysis and if it has a high p -values it not significant for the analysis and is dropped from the analysis.

D3

A screenshot of a computer

Description automatically generated

E1. So, after the data reduction with backward stepwise elimination we were able to reduce the model by 13 rows from 20 to 7. The analysis is relatively similar for both model with the reduced model having a higher log-likelihood score, but the original model having a slightly higher pseudo r-squared score, but both have a really low LLR p -value show these values are significant.

E2

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

E3. See code Main\_paper\_D208\_Logistic\_Regression.ipynb file atached

F1

F1A

|  |  |
| --- | --- |
| Ln(p / (1-p)) = -6.40 + 0.05(MonthlyCharge) – 0.10(Tenure) + 0.19(PaperlessBilling) + 0.28(Gender\_Male) – 3.08(Contract\_One\_Year) – 3.23(Contract\_Two\_Year) – 2.34(InternetService\_Fiber\_Optic) |  |

F1B. In this model we have found that that clients start off by being more likely to stay than not showing that the data is skewed toward no than yes and the intercept verifying this by starting off negative. We also found that having longer term contracts and having them stay longer helps to reduce customer churn also customer having Fiberoptic also helps with keeping customers. Other factors like paperless billing, increase monthly charge, being the gender male do lead to the customer churning, but not quite as much as having the as other factors benefit weather a customer does not churn.

F1C . I do find these finding significant the LLR-P-value is less than 0.00 showing these values are significant and running the accuracy scores they are all around 0.8 with accuracy being 0.9 showing the model is great for predicting whether a customer churns or not. I think this model is practical significant and would be good for predicting customer churn, but it is always good to test with other data to see if the model holds.

F1D. The Limitation with the dataset the original model seems to be slightly better so that is a limitation of the second model, but will run a lot faster because of data reduction. Another limitation is that I ran a backward feature elimination instead of RFE which might be a better model for predicting result especially when you introduce new data. I also did not run a cross validation model which might be better in seeing how other datasets might affect the analysis.

F2. My overall recommendation is to introduce a new data set to verify the accuracy of the model. I would conclude if this model holds up against other datasets, I would recommend the company to promotes longer term contract and fiber optics because they increase the chance a customer stays by quite a bit. To keep a look out for how much the company charges people because the more you charge the more likely the customer churns. To keep an eye out on features like paperless billing or customer service to males to maybe see why those factors might impact why a customer churns.