Using Machine Learning to Predict Telecom Churn

Machine learning is a powerful tool to use in today’s business world. It is used in all forms of industries for a wide range of tasks. It truly allows for a more intelligent approach at predicting the future. In this example we show how we can take Telecom’s data to predict customer churn.

# Data Wrangling

The first step of the data science process is data wrangling. For Telecom we start by loading our dataset using the pandas package.

Text

Description automatically generated with medium confidence

Looking at the top 5 rows to get an initial look at out data we can see there are 21 different columns. An important thing to look at is the types of data these columns are. We will using dtypes for this step.

A picture containing text, receipt, screenshot

Description automatically generatedHere we notice that that most of the columns are of an object type. There are a few interesting things to point out here. There is both a “Loyalty ID” and a “Customer ID”. These columns will require us too look into it to see how they relate. At the bottom you can also notice that “Monthly Charges” and “Total Charges” are different types. We will have to make these two columns of the same type. Another thing to keep in mind is that Tenure is an integer type. We will need to decide if this is something we want to keep this way or change.

To explore the “Loyalty Id” and “Customer Id” we will check and see if every row has unique Loyalty and Customer ID. A quick look at this reveals every row has a unique “Customer ID” but not every row has a unique “Loyalty ID”. Below we take a look at this duplicate “Loyalty IDs”.

Text

Description automatically generated

Next we continue by looking for any Null Items and there are none. Looking at the unique values for the different columns we notice that they are not binary categories but rather three. They are “No”, “Yes”, and “No internet service”. For the “No internet Service” let’s make sure that the “Internet service” is truly “No”.

A picture containing graphical user interface

Description automatically generated

We will do this for each column making sure the data is correct.

# Exploratory Data Analysis With Tableau

For this project we will use a powerful data visualization tool, Tableau, for the exploratory data analysis (EDA). Tableau has a neat feature where you can format the slides within in a story board style. This allows has a neat feature where you can format the slides within in a story board style. This allows a way for conveying the EDA in an organized manner. We will show an example of the EDA that was performed on Telco with Tableau below.

Chart, bar chart

Description automatically generated

We start off by looking at the churn by contract type. It is clear that month-to-month contracts are churning at the highest rate.

Next, we take a look at the contract type with payment method. Looking at the contract type along with payment method, Electronic Check customers are churning at a much higher rate of almost 55% compared to the other categories.

Chart, bar chart

Description automatically generated

An initial glance at churn and monthly charges shows there might be slight skew to right with the higher monthly charges. This is something we should investigate further.

Chart, histogram

Description automatically generated

Noticing that month-to-month seem to be churning more, I looked at the average monthly charges by tenure. There was no significance here lower tenured customers and monthly charges.

Chart, bar chart

Description automatically generated

However, looking at the churn by tenure seems shows an obvious pattern.

Chart, histogram

Description automatically generated

Chart

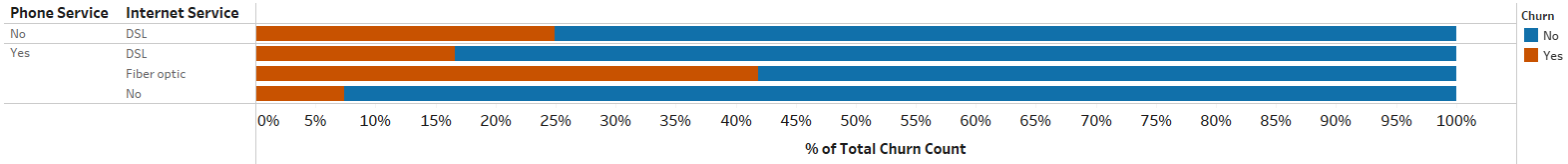
Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedI wanted to dig deeper into churn in relation to tenure. Below we show churn for customers with less than 5-month tenure. The relationship here shows that although the churn rate is almost 50/50, customers churning have a much higher average monthly charge. The question then comes to mind if this pattern persists with different tenure lengths. Below is the same charts with different lengths. The first shows tenure between 5-10 months and the second is tenure 10-15 months.

(From Left to right, Churn <5 months, Churn 5-10, Churn 10-15 months.)

After this we started to look at the churn by service types. Notice how customers with Phone Service and Fiber Optic Internet Service have the highest churn.



To expand on this it is a good idea to add in more services to see look at more detail.

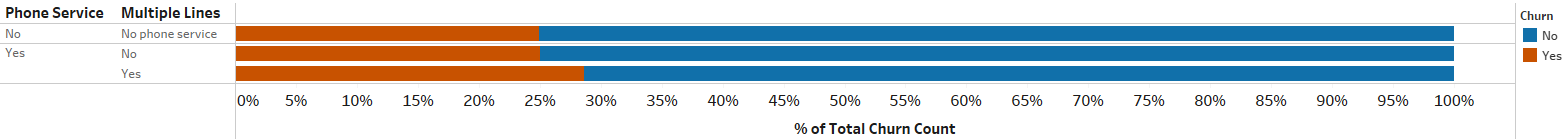
Chart, bar chart

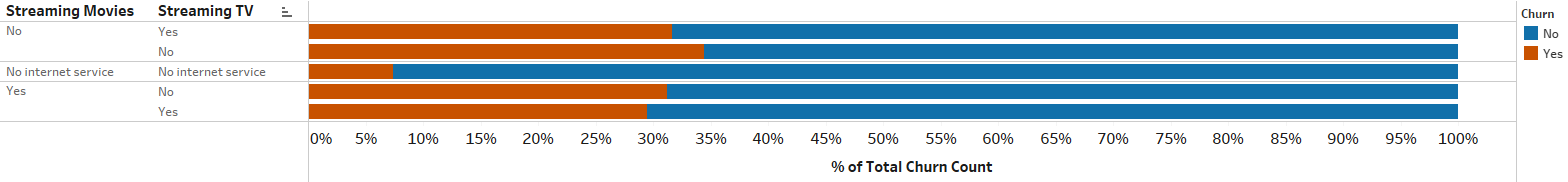
Description automatically generated

Clearly, Fiber Optic Internet has the highest churn ratio again. If the customer had Fiber Optic with no Device Protection, No Online Backup, and No Online Security churned at the highest rate. There is a pattern here where Fiber Optic Customers are churning more. This is important to note. Another thing here was the pattern with no Online backup or Online Security. Let’s explore that deeper.

Chart, bar chart

Description automatically generated

Taking a look at phone service we can see that there was no obvious pattern.

Streaming services also did not display any patterns.

Time to switch gears a bit and look at the customers by type. The first relationship we will look at is customers and whether they had multiple lines. Again here we saw no obvious patterns. Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedThe final thing to look into here is the relationship between churn and family types. Loking first if having dependents has an effect on churn and we can see there is a pattern.

Noticing this we add dependents. We notice on the top chart below that senior citizens with no family (partner or dependents) have the highest churn rate. On a different chart, on the bottom, we also add the multiple line descriptor. This does not immediately show any new patterns but could be something to look into further with our modeling.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

EDA through Tableau is a powerful tool to put together visuals and describe the thought process to the audience. It creates a great layout to take confusing datasets and visually show patterns that could be investigated later with a machine learning model.

# Preprocessing and Model

As far as preprocessing there is not much to do here. This dataset was clean. There were no null values. Therefore, to start off we will create a new data frame where we drop the the ID columns and the numerical columns that are not descriptors. This new data frame will be used to get initial scores for a chi square test of independence.

To start off we first look at the independence for Contract and type and Churn. This test looks at 2 separate variables and tests whether they are independent of each other or not. If the P-value is less than 0.05 than it rejects the hypothesis and signifies that they are no independent of each other. To set this up we use pandas crosstab to create a matrix between the two variables in question and then run chi-square contingency test.

Graphical user interface, text, application, Word

Description automatically generated

If we get a P-value less than .05 then it is significant and influences churn. In the above example we see that Contract type is a factor in regards to churn. We perform this same test on the variables that seemed to show a pattern in our EDA. For any the chi square tests where multiple variables were used we can create a new column in the data frame and combine these columns into one. Take this new column and use it to perform the contingency test. We can see below when we use the family columns that there is significance in terms of independence. This follows what we would expect from our EDA.

Graphical user interface, application, Word

Description automatically generated

To look at Tenure we do a simple correlation between churn and tenure.

Graphical user interface, text, application

Description automatically generated

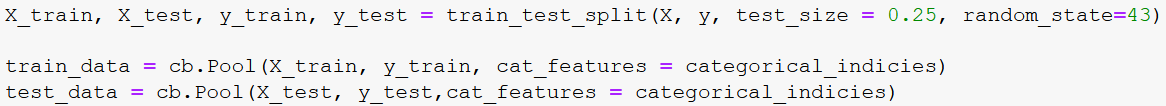
After running these stats we will then move on to the modeling section. We have to first set up the database and drop the ID columns. These will not be necessary for the modeling section. Something to fix in this new data base is the Total Charges column is blank when the customer is new (has tenure of 0). We need to fix this by filling in a zero. After, we need to change the Monthly and Total Charges to an integer type for the models.

We can then start turning all the categorical columns into category type. We do this by defining a function. Essentially what is happening is we are looking through the columns and choosing out the object type ones. We then extract the index for these and append them to a list. This list is used later one to input into the model.

Text

Description automatically generated

The CatBoost Model for classifiers will be used. Start of by splitting the data into a training and testing dataset. Notice how the categorical indices are used here. We will use a CatBoostClassifier and get the best parameters to predict.

  
Text

Description automatically generated

Table

Description automatically generated

Accuracy using this model so far is at 80% this is a decent score but ideally we should look into raising this score. After, some tuning the model was only raised to 81%. We could stay with this or continue to tune the model. This is time dependent on the user.

Using the get\_feature\_importance of CatBoost we can plot and show the features that are most important to driving the model. Below we see “Monthly Charges: and “Tenure” are the most important. Notice how this was seen in the EDA as well.

Chart, bar chart

Description automatically generated

Thinking back to the EDA a strong pattern shown was “Tech Support” and churn. Running a CatBoost model for this delivered a lesser accuracy of 73%.

The next model we use is a lightGBM model. This is a form of a gradient boosting model. We set this up with a parameter grid first. The reason behind the parameter grid is so that we can run a grid search and return the most optimal parameters for the model.

Text

Description automatically generated

These are only some examples of parameters that you can tune. Once this is ran and you get the best params keep running this model until performance stops improving. This can vary in length according to the data. One final parameter that was added in this example to improve performance was the scale\_pos\_weight parameter. This improved the recall score significantly. Below is the formula used to calculate this.



The results from the light GBM model produced are:

A picture containing calendar

Description automatically generated

# Future Work

Machine Learning provides a strong tool for future predictions. In our Telco example we ran through one model of prediction and achieved an accuracy of 81%. With further tuning or even a different model this could be improved. Creating a machine learning model is like creating a story and there are many possibilities.