Predicting Churn for Maximizing Profits

Tam Nguyen

January 2019

1 Introduction

1.1 The telecommunication industry

The telecom industry consists of internet providers and telecommunication services, and it one of the largest industries in the information society. The traditional source of revenue for telecommunication industry has been callings, but it's changing to be texting, image and video processing due to the decreasing cost of accessing internet.

There are some characteristics of the telecom industry:

- It has heavy regulations by the governmental and authorities (why)
- The market is highly competitive with different providers trying to provide the same service. Therefore, prices tend to decrease over time.
- Because there are large number of users national wide for telecom and internet services, the amount of data gathered from customers by these providers is huge. Knowing how to extract valuable information from the customers can help telecom companies gain competitive edges against other competitors in the industry.

1.2 Churn

Customer churn is one of the main issues in fields such as telecommunications, internet service providers, e-commerce, marketing, or banking. In the telecomunication industry, the market is saturated, making it very difficult to attract new customers. On the customer side, it is very easy to switch to a new provider if it provides a more beneficial offer (Canale and Lunardon, 2014). Because of this saturated market, the cost of aquiring new customers can be five to six times higher than retaining existing customers. Therefore, companies might be best invested in developing customers' trust on the service rather than attracting new customers.

Researches on predicting customer churn have been done to try to predict whether a customer is likely to quit the service of their service provider and join a different one. Having an understanding of the influential factors causing customer churns on the service can help companies understand customers' needs and adjust on their provided services based on these factors to reduce churn.

1.2.1 Types of churn

There are two types of churn.

- Voluntary churn: customers decide to leave the service and turn to another provider. Knowing why these types of churners decide to leave is critical for churn management
- Involuntary churn: there might be technical problems causing companies to discontinue the service itself.

Voluntary churn will be the one we are interested in knowing more. It is essential for any business to be aware of who is about to churn, when is the customer about to churn, why does a customer churn.

2 Method

This project uses R to implement three predictive models: logistic regression, random forest and decision trees. I choose these models because they are often used in papers that have been published on churn. These are also well-known predictive models.

Some of them will be implemented as black box, meaning I cannot know very well what are the processes they do the analysis on. Therefore, I will choose the dive deep into logistic regression and see if they are well performed. I choose logistic regression particularly because it is one of the models easiest to explain and measure performance. Bear in mind that these models will be briefly explained instead of looking into details. Specifically, I will mention its advantages and disadvantages of each of these models, then I will explain the steps I did to get the final results.

CRISP helps to build a data science project based on business results and financial benefits. I will follow this framework since it combines both the data science approach and the evaluation of the project. It connects the models with the return on investment to show the value or impact from the business.

The reasons we also use CRISP is because:

- It's an agile method: it implements data science projects iteratively, helping us to overcome common sense thinking and help us to adjust to unexpected issues quickly.

- It's data science for business's sake. It focuses on the ROI of the project, helping us to evaluate KPIs and potential economic impact if the model is implemented.
- In terms of churn, ROI is one of the most critical metrics in which we want to know besides the technical aspect of data science.

Follow this framework step-by-step helps us to make sure we don't miss any important steps during the process. The important steps are described as followed:

- Business Understanding: In this step, we define the business objective of the project, identifying the appropriate KPIs and create a plan to correspond to the goal of the project
- Data Understanding: This step requires getting familiar with the data and collect the data in appropriate formats for analysis.
- Modeling: We apply predictive models and try to optimize the results.
- Evaluation: We use various methods to assess the model results, clarifying the black box models to make it understandable to different stakeholders. Some of the methods we use in this project are the ROC curve, gain and lift chart and LIME.
- Deployment: We use the results from the model to link with the business processes, providing the recommendations needed to improve the business.

3 Business Understanding

3.1 The description of the data set

Our data set consists of 7043 profiles of telecom customers and is available via.... The data contains of two main types of variables:

- Customers' personal characteristis such as their gender, whether they have a partner, their tenure status, whether they are senior citizens.
- There usage behaviors such as phone service, internet service, online security, tech support, streaming TV, streaming movies, contract, payment method and their month charges.

Both of these types of variables are beneficial for our predictive analysis. It helps us to detect whether their decisions to churn is based on personal characteristics or on their service usage behaviors.

	Churn	Count	Percentage
1	No	5174	0.73
2	Yes	1869	0.27

Table 1: Percentages of churners versus non-churners.

The table shows an imbalanced distribution between churn and non-churn. There are only 27% churners out of the whole dataset. The industry average of churn accross the US is about 1.9%. While this is a huge difference, the average churn rate in the US is assessing over millions of people, whereas in this dataset we only have 7043 observations.

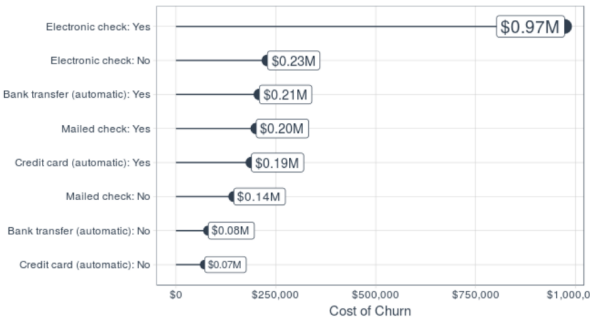

Figure 1: Estimated cost of Payment Method and Paperless Billing.

This figure is generated

4 Data Understanding

4.1 Exploratory Visualisation

Exploratory visualization helps us to gain a general understanding of the relationships between explanatory variables and the response variable. The first step in this visualization process is to evaluate the distribution of different variables corresponding to churn. Here we visualise a summary of the distribution of different variables with two charts. One is for numerical variables and another is for categorical variables.

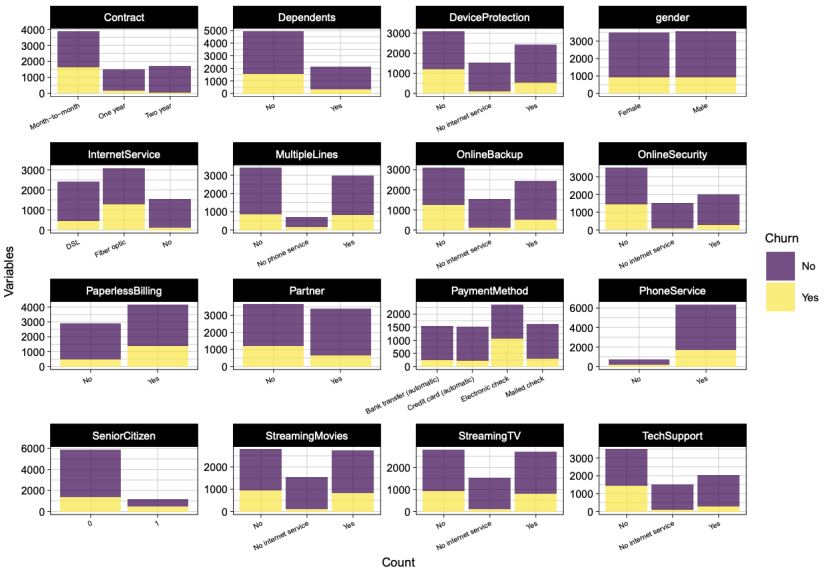

Figure 2: A distribution of numerical variables in respect to churn

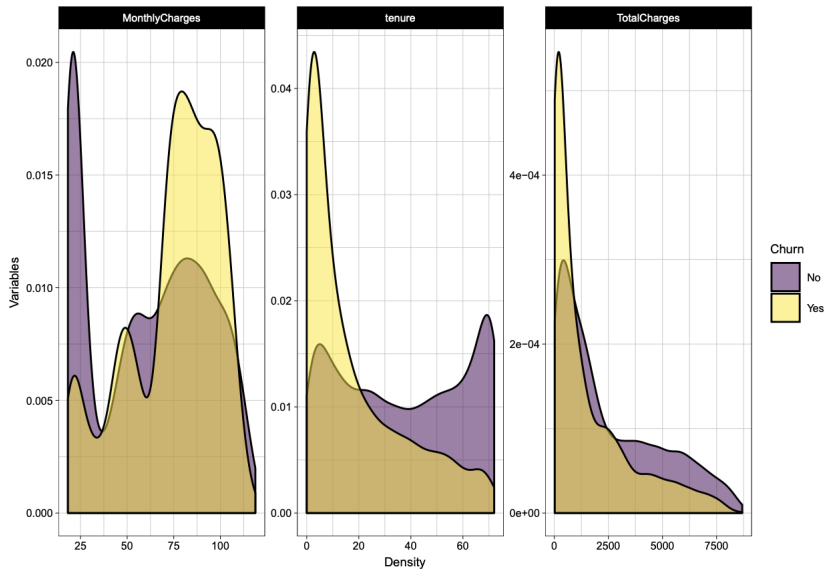

Figure 3: A distribution of categorical variables in respect to churn.

There are not a lot of visible evidence suggesting if churners show a lot of differences from non churners, especially if we draw from the graph showing only categorical variables. Whereas, for our numerical variables shown in figure 2, it suggests that churners have to pay higher monthly charges and total charges than non-churners and have a higher tenure rate earlier. From this graphs, we hypothesis that the monthly charges and tenure are two of the significant factors that affect the rate of churn.

4.2 Correlation Analysis

Correlation analysis shows linear relationship between variables. Specifically, it assesses how much one variable changes when the other changes. In this specific case, we want to assess the correlation between other explanatory variables and the response variables. Which explanatory variables have the most correlation with the response variable, which is churn, can be projected to give us some clues about the potential variables affecting churn. From there we could generate our hypothesis for the modeling process.

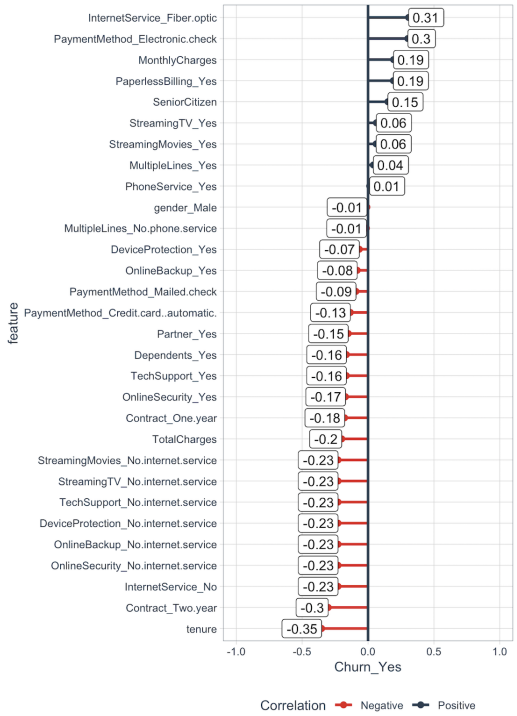

Figure 4: A correlation plot of all variables in the dataset in respect to churn.

This correlation plot shows which variables are correlated to churn and oth-

erwise. We found that Internet Service with Fiber optic, Electronic Payment Method, Monthly Charges, Paperless Billing and Senior Citizens all have more than 10% correlation rate with churn, compared to other factors. From this plot, we can further our hypothesis by paying more attention to these factors and see if our model's prediction resembles those variables in our correlation analysis.

5 Hypothesis Summary

From our previous phases, we have accumulated a number of hypotheses that can be used to test for our predictive model. Our hypotheses are as followed:

- Tenure and monthly charges are the significant variables affecting churn.
- Payment method and types of Internet Services are also critical variables.

6 Modeling

In this paper, we use H2O for our analysis. Particularly, we use H2O autoML which combines different algorithms and score the performance of each model using a leaderboard. The implementations and details of these models used are not discussed here in detail. Rather, we will introduce the model used once we completed the modeling process.

(leaderboard)

Ensembles is a type of machine learning methods that combine multiple algorithms. In this specific case, H2O is a supervised machine learning algorithm that finds the optimal combination among many prediction algorithms called stacking. Stacking is a way to combine multiple algorithms by applying models by the results of other models. From a learned model, a new model uses the result of this model to train new models. There are different types of models tackling different spaces of the problem. The final model is stacked on top of other models which tackle each part of the modeling process. This is said to improve the overall performance of the model. The algorithm for stacking ensemble is as followed:

- Split the training set into two sets, such as 0.3 and 0.7
- Train several base learners on the first part.
- Test the base learners on the second part.
- Using the predictions from the testing result as the inputs, and the correct responses as the outputs, train a higher level learner.

In stacking, the combining mechanism is that the output of the classifiers (Level 0 classifiers) will be used as training data for another classifier (Level 1 classifier) to approximate the same target function. Basically, you let the Level 1 classifier to figure out the combining mechanism.

7 Evaluation

7.1 ROC

We are using LIME, ROC, and gain and lift charts to measure our model performance. One of the most commonly used way to measure it is the ROC curve.

It is quite difficult to handle different expected calculations that we need a graph to evaluate it in different thresholds (visualize the performance possibilities of every threshold?). One such curve which helps us to visualize it is the Receiver Operating Characteristics graph, which shows the false positive rate on the x axis and the true positive rate on the y axis. This shows the relative tradeoffs between benefits and costs, or true positives versus false positives.

The tradeoff is followed: if we want high true positive rate, we have to expect a high false positive rate. This should be taken into consideration because in many real world examples, a moderate amount of false positives can cost far more than the other false negative. What about in the case of churn? In churn specifically, we found false negatives as the most costly classification among all other scenarios. While true positives help us to gain profits, false positives cost us an amount of money, but the most costly would be false negatives when we incorrectly identifying them as churners but actually not churn. We lose the most amount of money. It is important, therefore, to minimize our false negative rate while maximizing true positives rate.

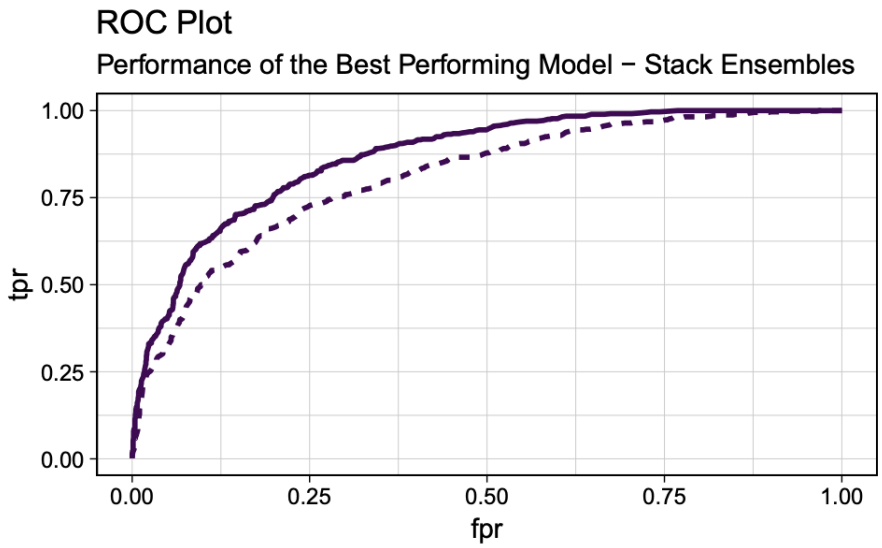

Figure 5: The ROC curve of the best performing model - H2O Stack Ensemble.

Our ROC curve shows a high performance of the H2O Stack Ensemble model. The model performs much better than a random average model. From this graph, we can calculate the summary statistics for this ROC curve and calculate the optimal threshold value to maximize profits.

7.2 Confusion Matrix

		Predicted	
		No	Yes
Actual	No	True Negatives Predicted Customers Stay Customers Actually Stay	False Positives Predicted Customers Churn Customers Actually Stay
	Yes	False Negative Predicted Customers Stay Customers Actually Churn	True Positives Predicted Customers Churn Customers Actually Churn

Table 2: Confusion Matrix Illustration. The red color in the table shows the most costly rate in predicting churn.

Table 1 is an illustration of different scenarios between predicted events and actual events. The cost of the model is derived from the false positive rate and the false negative rate. The most costly mistake of the model is resulted from the false negative rate, where we predicted customers stay while customers actually churn. The explanation of why it is the most costly will be illustrated in the expected value section in which we evaluate the costs and returns of each scenario.

		Predicted		Error rate
		No	Yes	
Actual	No	2624	437	0.14
	Yes	218	889	0.20
Totals		2842	1326	0.16

Table 3: Confusion Matrix of the Stack Ensemble model.

Table 2 the confusion matrix result from the Stack Ensemble model. From the table, we can derive that the false negatives rate is 437, which is 0.33% error rate out of the whole negative rates. The false positive rate equals 218, which is about 7% out of the whole positive rate. Thought Stack Ensemble is the best model, the error rate is still quite high since a large false negative rate can cost us more than other rates.

7.3 Gain and Lift

Lift Chart are used measure your prediction of random guess vs using a model. Such improvement of prediction from random guess is called Lift. Lift is a measure of the performance of a targeting model at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model.

Gain = (Expected Response Using Predictive Model) / (Expected Response From Random Mailing)

Lift = (Expected Response In A Specific Lot Of 10,000 Prospects Using Predictive Model) / (Expected Response In A Random Lot Of 10,000 Prospects Without Using Predictive Model)

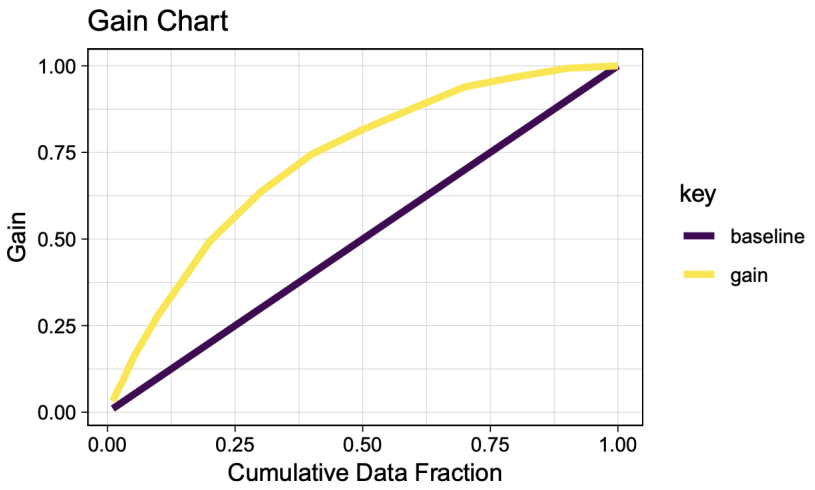

Figure 6: Gain chart of the Stack Ensemble model.

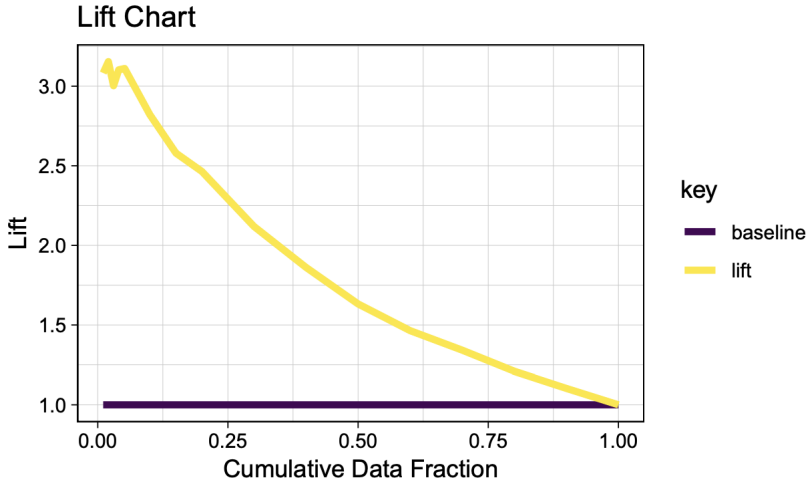

Figure 7: Lift chart of the Stack Ensemble model..

Gain and Lift description goes here

7.4 LIME

The common problem of any complex model is that it is a black box model and is extremely hard to interpret. So it is quite difficult to trust the model. LIME

provides a local interpretation, estimating which feature adds the most value to the prediction. Because it makes complex models interpretable, it is good for human practitioners, businesses to understand and build trust to the algorithm.

Variable importance only shows how correlated the relationship between the variable predicted and the response variable. It does not give us whether given a new observation, what are the most influential predictors influencing the outcome. Though a most important variable might not drive the reason causing churn. We need to understand what is most influential for the specific employee.

Local Interpretable Model-agnostic Explanations (LIME) is a visualization technique that helps explain individual predictions. As the name implies, it is model agnostic so it can be applied to any supervised regression or classification model. Behind the workings of LIME lies the assumption that every complex model is linear on a local scale and asserting that it is possible to fit a simple model around a single observation that will mimic how the global model behaves at that locality.

The general algorithm of how LIME works is:

1. Given an observation, permute it to create replicated feature data with slight value modifications.
2. Compute similarity distance measure between original observation and permuted observations.
3. Apply selected machine learning model to predict outcomes of permuted data.
4. Select m number of features to best describe predicted outcomes.
5. Fit a simple model to the permuted data, explaining the complex model outcome with m features from the permuted data weighted by its similarity to the original observation .
6. Use the resulting feature weights to explain local behavior.

different probability to churn and best explain the linear model in the local region. The plot also provides the model fit in the "Explaination Fit" description, which explains how well the model explains the local region.

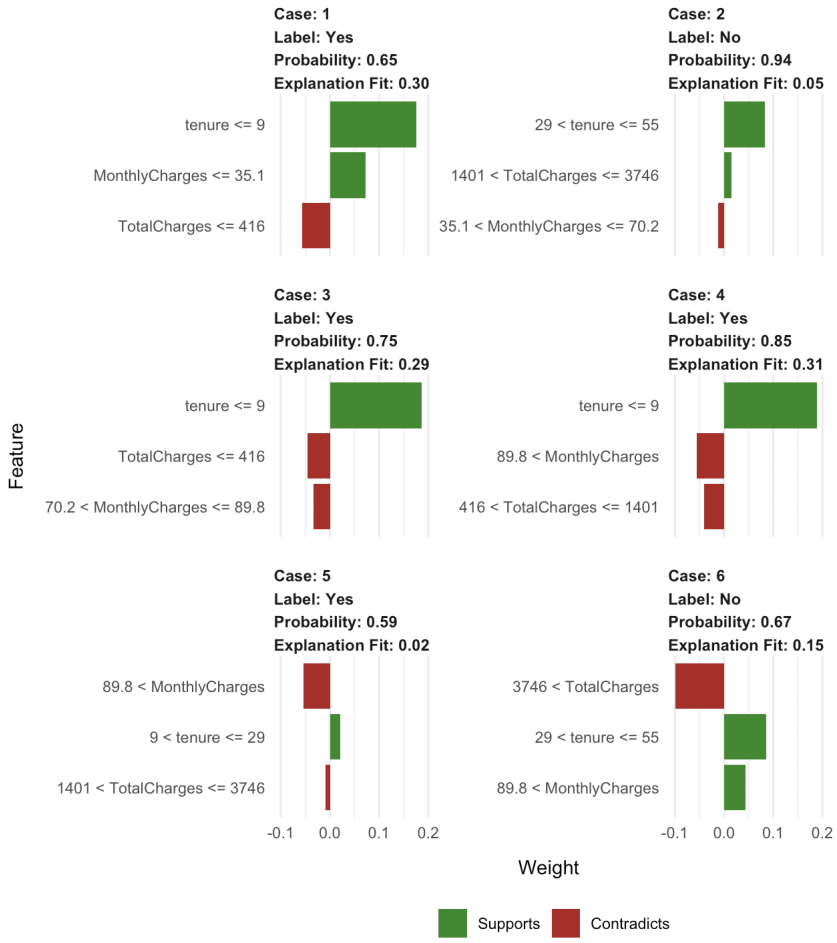

Figure 8: A Multiple Explanation Plot based on LIME framework.

From the graph, the three variables repeated in this LIME plot is tenure, monthly charges and total charges. The patterns in the plot show that early tenure is the highest factor causing people to churn.

(we're not sure if we can trust the model since they show high inconsistencies in total charges and montly charges)

The second LIME plot shows a heatmap with variables of different influence. This is useful for seeing common features that influence all observations. This also shows the feature weight of variables corresponding to churn.

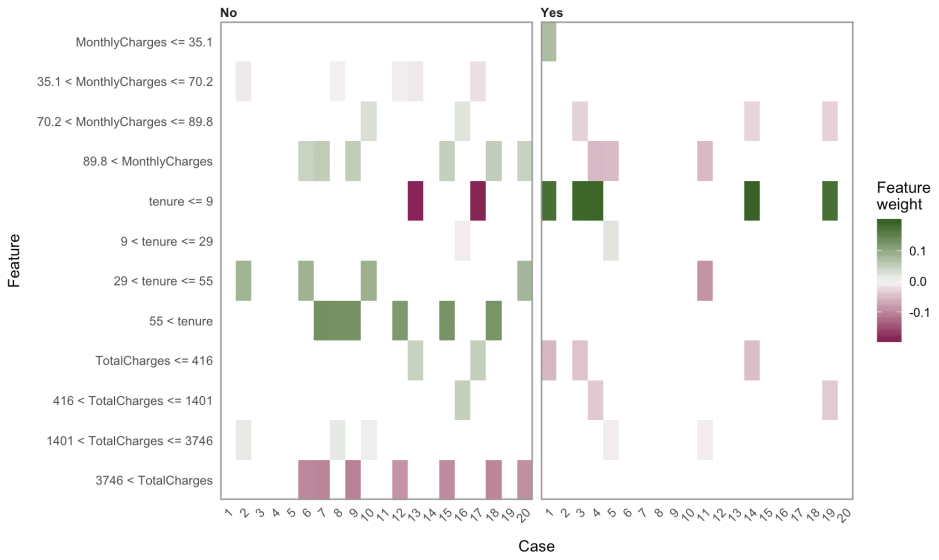

Figure 9: A Heatmap of all variables in respect to churn based on LIME framework.

From this graph, only tenure smaller than 9 shows a high feature weight compared to other variables. The higher the tenure, the feature weight corresponds to non-churn increases.

Higher total charges and monthly charges seem to show higher feature weight with non-churn. This does not adhere to our hypothesis and what the evidence in our correlation analysis shows.

8 Expected Value Framework

	Cost	Revenue	Profit
True Positives	15% discount for 33 months (74.41-63.25)*33 = 368.28	Earn 33 months of revenue 63.25*33 = 2087$	1718.72$
True Negatives	If a customer does not churn, it has no effect on our model 0	0	
False Positives	15% discount for 33 months (74.41-63.25)*33 = 368.28		-368.28
False Negatives		Lose 33 months of revenue 63.25*33 = 2087$	-2087

Figure 10: A boat.

8.1 Threshold Optimization

9 Conclusion