Predicting Churn for Maximizing Profits

Tam Nguyen

January 2019

1 Introduction

1.1 The telecommunication industry

The telecommunication industry consists of internet providers and telecommunication services. In 2017, First Research, a market research form, estimated revenues in the US telecommunication industry at 590$ billion dollars, making 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. The lowering cost also allows people to connect and subscribe to telecommunication services. Many Western countries have their phone penetration rate of over 100%, meaning there are more subscribers than citizens [1].

Though this is a big industry, most telecommunication companies in developed countries are struggling to maintain profitability [2]. The market is almost saturated with high competitions since different providers sell the same services (i.e wireless, phone plans). Often the fights are between very large companies, each has about 30% of the market share [3]. Small and midsize businesses compete with large companies by providing services in specific regional coverage. However, larger companies tend to create partnerships with these small companies for further acquisitions and expansions [4].

Because there are large number of users national wide for telecommunication 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 telecommunication 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. Because of this saturated market, the cost of acquiring new customers can be 50 times higher than keeping a customer, in terms of wireless subscription, and it is still increasing over time [2] [3]. 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). Therefore, companies might be best invested in developing customers' trust on the service rather than attracting new customers.

However, though top executives in companies report customer retention (the reverse of churn) as well as reducing churn is one of their main objectives, 49% of them reported being not satisfied with the ability to support their companies' retention goals. From the customers' points of view, 85% customers report companies could do more things to retain them. Many studies also found the reverse benefits of retention campaigns as they are ineffective keeping their customers [5]. This shows that there are still gaps that could be addressed in cases of churn and that executives often underestimate the financial impact churn can cause.

Additionally, managing customers to reduce churn and retain them have been shown in literature to be profitable to companies because (1) they spend less money in acquiring new customers, (2) long-term customers tend to generate higher profits, less costly to serve, and may provide new referrals through positive words of mouth (New insights). On the other hand, losing customers can lead to (1) negative words of mouth, (2) more opportunity cost because of reduced sales [6]. Often, in terms of acquiring new customers, the acquisition cost in the telecommunication industry is very high due to fierce competition. In Canada, Bell and Telus reported an average customer acquisition cost of 521$ while the retention cost for each subscriber is about 11$ [3]. Therefore, small improvements in reducing churn can generate a significant increase in profits and decrease in costs. A report by McKinsey estimated that reducing churn could increase the earnings of a typical wireless carrier in the US by about 9.9% (Braff, Passmore..).

Researches on predicting customer churn have been conducted 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 categorized by Larazov and Capota (2007):

- Voluntary churn: customers intentionally 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 is critical to businesses because it addresses what companies could change or further optimize to minimize churn. To make changes, business practitioners should take care of three main dimensions: *who* is about to churn, *when* is the customer about to churn, and *why* does a customer churn [**?**].

1.3 Approaches to churn

Based on various studies on predicting churn, the methods used in these studies range from RFM (recency, frequency, monetary) statistical models to ensemble learning models such as random forest. Studies often found that machine learning methods outperform traditional statistical methods [7]. The best model in terms of predictive power and understandability (i.e. how easy it is to interpret the result of the model) is Decision Trees [6].

A lot of debates have moved beyond the selection of models to questions of why customers are at risk, whether they can be retained and what incentives companies could give them to increase retention [7]. Some researches also give new insight into the social network model of churn, claiming that customers are less likely to churn if their contacts also use and increase the number of usage of the service.

Researchers also discuss how to best target customers to maximize profits. There are differences between customers who have high probability to churn versus customers who should be targeted. Aurelie and Sunil also find problems with the existing practices to predict churn. The traditional approach is coming up with a model to predict the percentage of churn for each customer, then select the top few percents of those who are likely to churn and offer them incentives to stay. However, this approach is not optimal for finding the most profitable customers and who are most likely to respond when we offer incentives. New profit-driven methods are being introduced by recent researchers to further optimize churn prediction.

Another technical challenge found in churn is the imbalance structure of the data, since the number of churners are always much less than that of non-churners. There are different methods used to make the dataset suitable for the modeling process such as noise removal, undersampling, normalization and feature selection. For the imbalance data problem, undersampling has been used as a technique to balance the number of churners and non-churners. However, it has been found that during this process a lot of useful information was lost. All of these debates about what best to predict customers and retain them are still going on and there are still a lot of potential to find more optimal solutions for

this churn prediction issue, let alone the practicality of these researches. While a lot of researches have proposed ideas to tackle churn in terms of theory, they have not yet provided how to implement them.

1.4 Rationale of this study

This study aims to predict churn using Automated Machine Learning (Auto ML) and measure the expected value of the model to maximize business value for the telecommunication industry. The entire project uses R as a programming language to conduct data preparation, modeling, evaluation. I also aim to communicate results effectively using data visualization, as it is usually the best way to understand and gain insights.

Though many papers on churn review existing approaches such as what models are often used, this paper focuses more about communication, data visualization and evaluating business values of the model. The details of the technical aspects, such as the mathematical aspects of the machine learning models, or the code used to implement these models and visualizations are not the main focus of this project. However, for reproducibility, one can access the code of this project via the link and implement it on their own via RStudio Cloud.

2 Method

The project follows the Cross-industry standard process for data mining framework (CRISP). This framework 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.

Figure 1: Illustration of the CRISP framework.

The reasons we also use CRISP is because it is It's an . Ile method: it implements data science projects iteratively, treating each process as a circle, coming back to the previous processes if the results were not optimized. There are other processes focusing more on the analysis part of the project and less on the business aspect. In this project specifically, we focus on the business aspect and the potential ROI when the model is implemented. In terms of churn, ROI is one of the most critical metrics in which we want to know besides aspects of data science.

- 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. Also, exploratory visualisation is often used to detect patterns of the data and generate hypothesis.
- Modeling: We apply predictive models and try to optimize the results. In this project, we will use H2O.ai framework for our modeling process. This allows us to focus on model's performance and less on the technical aspect.
- Evaluation: We use various methods to assess the model results, clarifying the black box models to make it understandable to different stakeholders (i.e. black box models here mean a model too complex that it is not clear how each input contributes to the output). Some of the methods we use in this project are the confusion matrix, the ROC curve, gain and lift charts and the Local Interpretable Model-agnostic Explanations (LIME).
- Deployment: We use the results from the model to link with the business processes, providing the recommendations needed to improve the business.

In this particular step, this project will evaluate the expected value of the model given and calculate the potential profit gained if implemented.

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 characteristics 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.

In churn prediction cases, studies often found attributes such as customer bills, call duration details, customer demographics, age's of customer smart phone and the average cost as important predictors of churn. In this dataset, we do not have call duration details and very few details on customer demographics such as where they live or what are their income profiles. Therefore, we cannot test all of these variables, and can only focus on some variables relevant to the dataset.

Despite of these limited number of attributes, 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.

Hypothesize 2 important variables and make a calculation like this.

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 across 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.

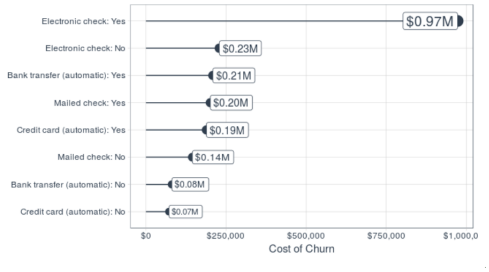

Figure 2: Estimated cost of Payment Method and Paperless Billing.

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 visualize a summary of the distribution of different variables with two charts. One is for numerical variables and another is for categorical variables.

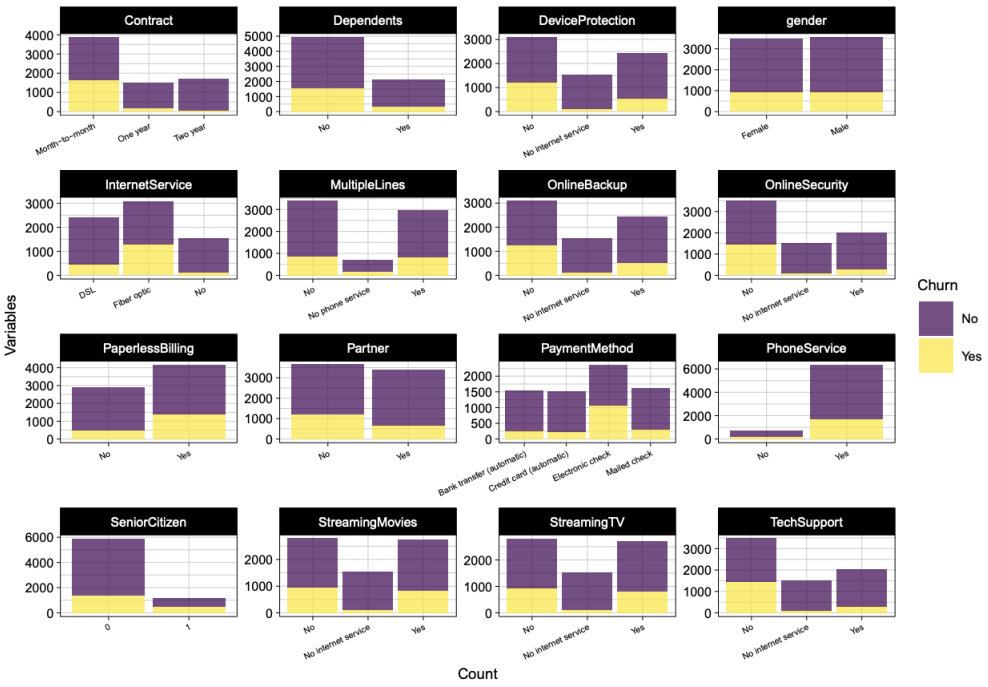

Figure 3: A distribution of numerical variables in respect to churn

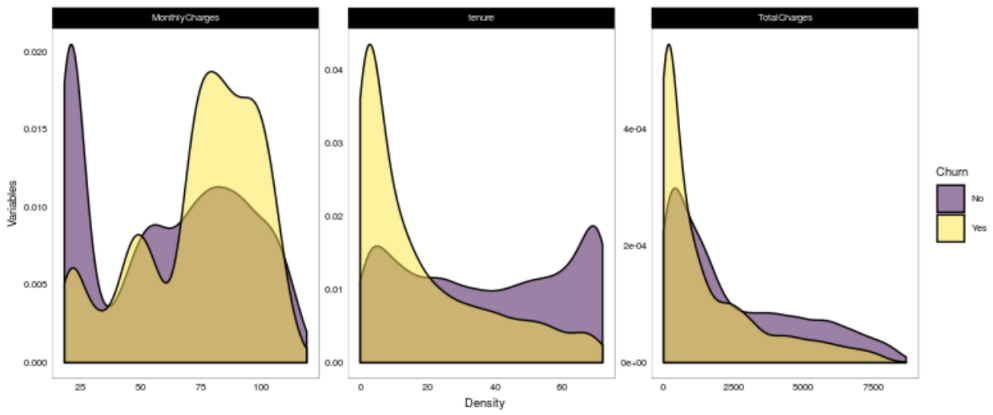

Figure 4: A distribution of categorical variables in respect to churn.

From these two graphs, the first grasp is the distribution of rThere 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.

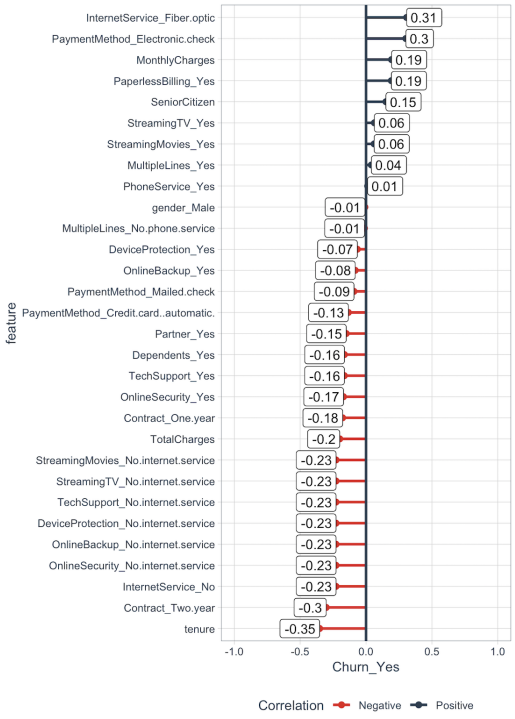

Figure 5: A correlation plot of all variables in the dataset in respect to churn.

This correlation plot shows which variables are correlated to churn and otherwise. 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

6.1 Data preparation

Before using models to predict our churn data, some data preparation and variable selections techniques need to be implemented for (1) the optimization of analytical models and (2) removing messy data that will cause bad performing models. I first treat missing values since they do not contribute to the model. The method proposed by Verkebe (2012) is applied. If more than 5% of observations are missing, then some imputation techniques were applied since the number of observations can have an impact on the model. If missing values are less than 5%, it is better to remove them since the overall number of removed instances remained not significant. For the current dataset, there are 17 (check) observations missing, making is insignificant to impute them. The missing values are then removed for the modeling process.

The only preparation step being implemented is removing variables that have only one single value. Another variable being removed is CustomerID, since it is an unique number ID of a customer and has no contribution to the modeling process. As a result, no other variables except Customer ID are being removed, since there is no variable in the dataset that has zero variance (or have only one single value). Churn is set in our modeling framework as the response variable. Other data preparation steps, if necessary, are handled by the framework being introduced shortly.

6.2 Modeling using H2O

In this paper, H2O.ai framework is used for the modeling process. H2O is an open-source machine learning framework, which provides us the ability to im-

plement machine learning algorithms without investing in much of the details and technicalities. Particularly, we use H2O AutoML, which aims to automate the machine learning workflow and train different algorithms within a specific time-frame. The performance of these implemented models is shown in a leaderboard. The implementations and details of these models used will be introduced once we completed the modeling process.

Using the programming language R, the entire dataset is split into two different datasets, one is for training and the other is for testing. The split up is 70% for the training set and 30% for the testing set. H2O AutoML is applied to train with maximum 7 models. The maximum training time is set to 10 minutes. Once completed, the leaderboard showing the performance of those models is shown in the graphic below.

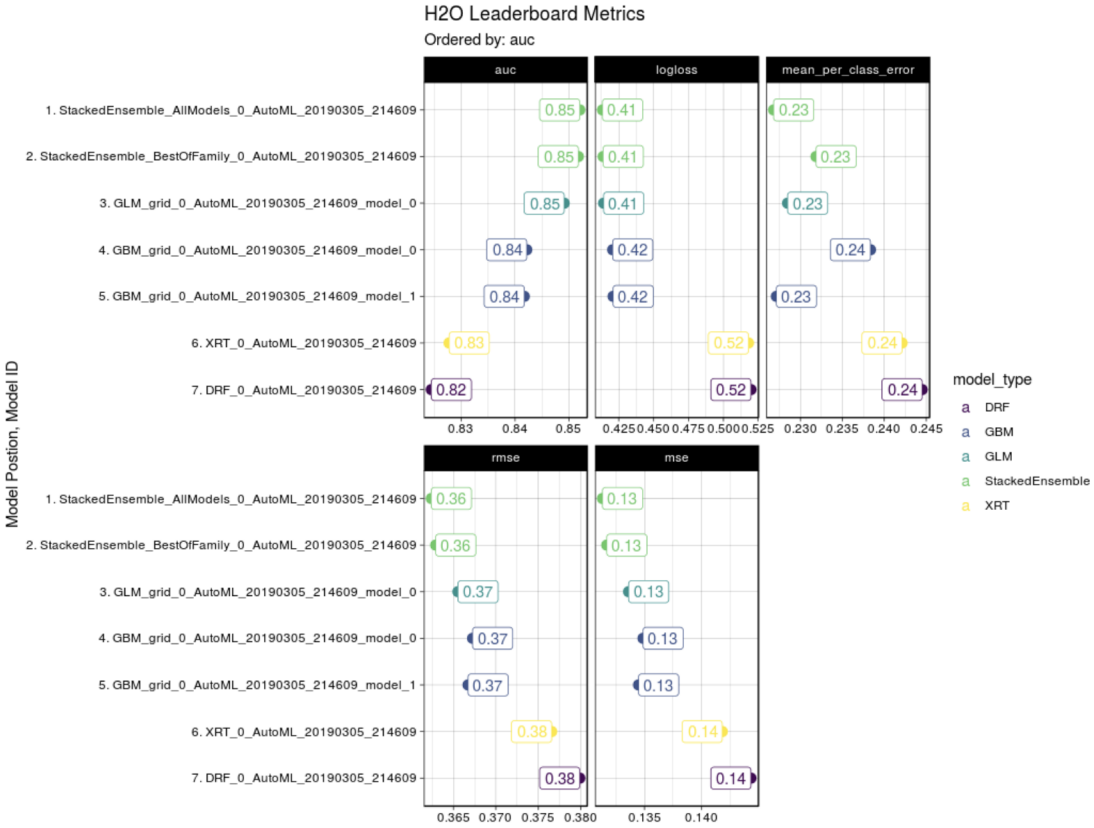

Figure 6: The leaderboard of the models implemented using H2O AutoML ordered by AUC.

The leaderboard is ordered by AUC, which is the Area under the Receiver Operating Characteristics Curve, a metric that is used to measure the performance

of a model. Intuitively, it provides an estimate of the probability of a random set of observations of churners is ranked correctly and higher than the selected observations in the class of non-churners. In other words, it is the probability that a churner is assigned a higher probability to churn compared to that of a non-churn. In figure 5, the best performing model has an AUC of 0.85, and has a smaller mean per class error. Please note that on the graph the metric of mean per class error shows 0.23, which is the same. However, the difference is not shown since 0.23 is rounded. In further decimal points, the differences are shown.

Ensemble Stacking is ranked as the best performing model. There are more than one types of the same algorithm (one is StackEnsemble with All Models versus Stack Ensemble Best of Family). Other algorithms being implemented and ranked on the top are default Random Forest (DRF), an Extremely Randomized Forest (XRT), and three pre-specified XGBoost GBM (Gradient Boosting Machine) models. Stack Ensemble at the first place is not a surprising result since it is based on all previously trained model and is trained on the collection of individual models to produce a high predictive power. It is one of the most powerful algorithms in this H2O framework.

To explain how Stack Ensembles works in detail, it is a type of machine learning methods that combine multiple algorithms. It 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 (i.e classifiers, models) 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 previously trained models will be used as training data for another model. The entire process resembles a voting procedure to obtain a final prediction.

From this point, the leader of the trained models, Stack Ensemble, will be used in the rest of the paper. One drawback of the Stack Ensemble algorithm is that it is implemented as black box models, making it difficult to interpret and understand. Because of this, a large number of researches in churn prediction have been using Decision Trees and rated as one of the classification technique

that yield significant predictive power and has a high comprehensibility. In this comprehensibility aspect, later in our evaluation part, we tackle this specific issue by applying a framework that tries to solve this problem of black-box models called LIME (Local Interpretable Model-agnostic Explanations).

7 Evaluation

7.1 Confusion Matrix

Once the modeling process is completed, several metrics are derived to assess the performance of the model. The first metric is the confusion matrix, which is a two by two matrix categorizing the number of correct and incorrect predictions for both positive and negative class. Positives means customers being classified as churn, and negatives means customers classified as staying. The true negatives and true positives are the correctly classified observations for the positive and negative classes. Likewise, the false negatives and the false positives represent the errors the model incorrectly classified. Usually, incorrectly predicted observations have different importance, meaning false negatives can cost more than false negatives.

		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.

In our illustration shown in Table 2, the highlighted box shows the importance of false negatives, the most costly classification error of the model. This is when the model predicts customers stay while customers actually churn. In other words, when the model predicts customers stay, there is no need to give discount incentives for these customers. The company thereby loses profits by making them leave without doing any action for them to stay. Whereas, in terms of false positives (the model predicts customers churn while they stay), the company reduces profits by offering discounts, but still have some revenue

from these returning customers. It is important, therefore, to minimize our false negative rate (or minimize costs) while maximizing true positives rate (maximize profits).

		Predicted		Error rate
		No	Yes	
Actual	No	1287	263	0.17
	Yes	159	401	0.28
Totals		1446	664	0.20

Table 3: Confusion Matrix of the Stack Ensemble model.

Table 2 shows the confusion matrix results from the Stack Ensemble model. From the table, we can derive that the number of false negatives is 159, which consists of 10% of the negative rate. Likewise, the number of false positives is 263. The total error rate is 0.20. (explain this further)

7.2 ROC

The other most commonly used ways to measure model performance is the Receiver Operating Characteristic curve (ROC), which shows the false positive rate (customers that stay the model incorrectly identify as leaving) on the x axis and the true positive rate (customers the model correctly identify as leaving) on the y axis. The closer the curve reaching point 1 on the top left of the curve, the better the model is since we maximize correct predictions and minimize incorrect ones.

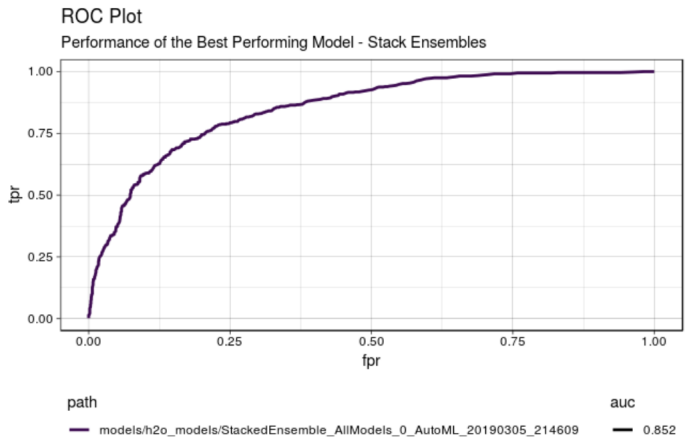

Figure 7: The ROC curve of the best performing model - H2O Stack Ensemble. The horizontal line shows the false positive rate and the vertical line shows the true positive rate.

Our ROC curve shows a high performance of the H2O Stack Ensemble model with the AUC score equals 0.852, as shown in the leaderboard. From most of the cases, if the AUC score is over 0.80, it is rated as a very good performing model. As shown in the graph, the model performs much better than a random average model, showing it has more true positives rate than false positives rate. The model, therefore, has a good performance overall. From this graph, we can calculate the summary statistics for this ROC curve and calculate the optimal threshold value to maximize profits.

7.3 Gain and Lift

Gain and Lift are important metrics especially for people who focus on the direct benefits of the model, one of the groups of stakeholders that might benefit from this is business people. The gain chart, specifically, measures what can be gained by using the model. As shown in the graph, the horizontal line represents the cumulative data fraction and the vertical line represents the gain. Based on the yellow gain line in the graph, we can interpret that targeting 37.5% of the high probability customers (cumulative data fraction) can potentially yield 75% of potential positive response. This is a potential result from the model, showing the model is performing well and it can be used to target customers with high benefits.

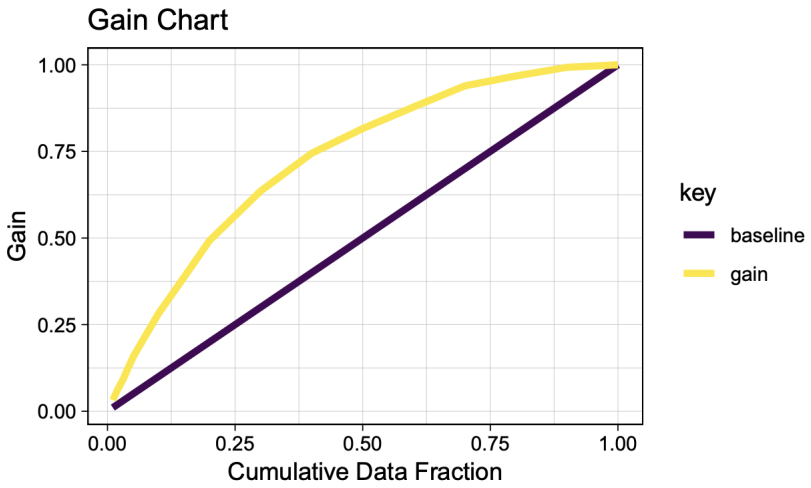

Figure 8: Gain chart of the Stack Ensemble model.

Lift Chart goes hand in hand with Gain Chart by showing the result of the modeling approach versus targeting people at random. Basically, lift is used to measure the prediction of random guess vs using a model. Such improvement of prediction from random guess is called Lift. It is shown that there is a direct

connection between profitability and lift (neslin 2006), thus, life has been used in many churn prediction studies as a performance criterion.

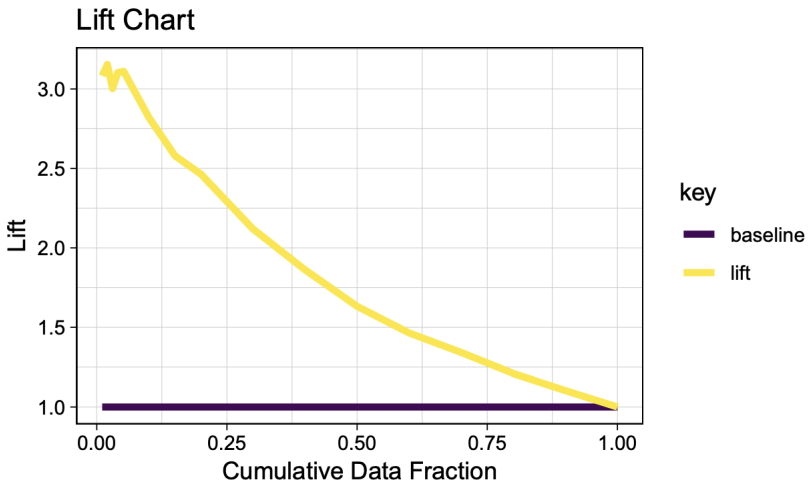

Figure 9: Lift chart of the Stack Ensemble model..

For this lift chart, it is shown that if we targeted 25% of people with high probability of churning people (cumulative data fraction), we have 2.5 times better targeting ability compared to targeting randomly (lift). This can potentially reduce cost by about 2.5 times versus random selection because we only need to offer discount incentives to customers with high probability to churn.

In various churn prediction studies, the average top decile lift is about 2.1 to 1, meaning customers in the top decile lift were 2.1 times likely to churn than average. Compared to our model's lift graph, the top decile lift is about 3.0 to 1. This shows the performance of Stack Ensemble is more accurate than average.

7.4 LIME

Usually in machine learning projects, the common problem of any complex model is that it is a black box model and is extremely hard to interpret. When we do not know how the inputs contribute to the output, it is difficult to trust the model. Local Interpretable Model-agnostic Explanations (LIME) is designed to tackle this particular problem. It 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.

LIME also has 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.

The LIME graph below shows different probability of churn for each case (i.e. observation) with the model fit in the "Explanation Fit" description, which explains how well the model explains the local region. There are 6 observations in total. For a larger number of observations, we will use a heat map, which will be introduced shortly.

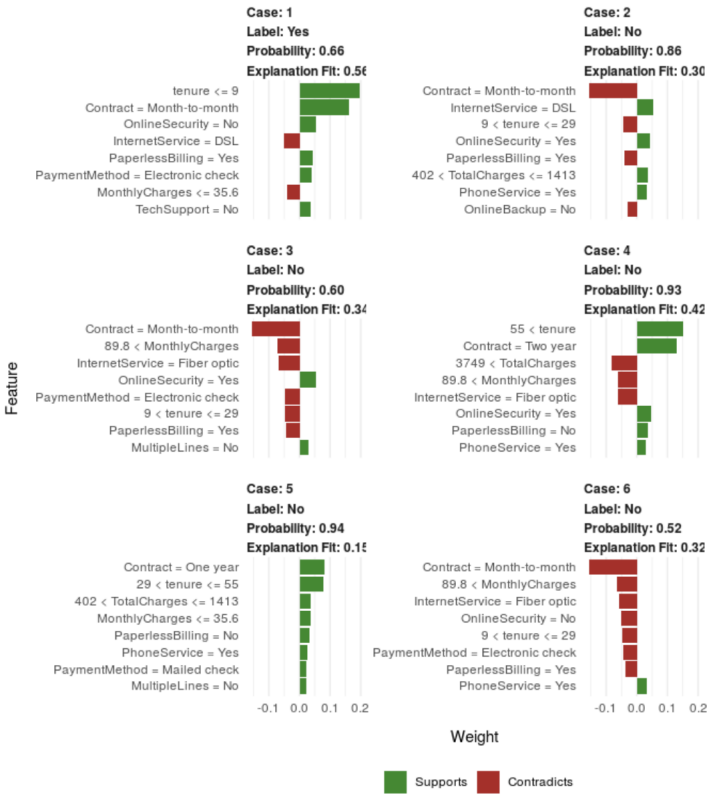

Figure 10: A Multiple Explanation Plot based on LIME framework.

From the graph, only case 1 shows the explanation fit of the person who churns, whereas the other cases are all predicted as non-churners. With non-churners, the tenure is often low with contract duration high. With contract duration low (e.g. one month), the feature weight usually contradicts with non-churn, suggesting that high contract duration correlates with non-churn. High monthly charges also contradict with non-churn so usually it is an indication of churn.

For a larger number of observations and clearer visualization, the second LIME plot shows a heat map 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.

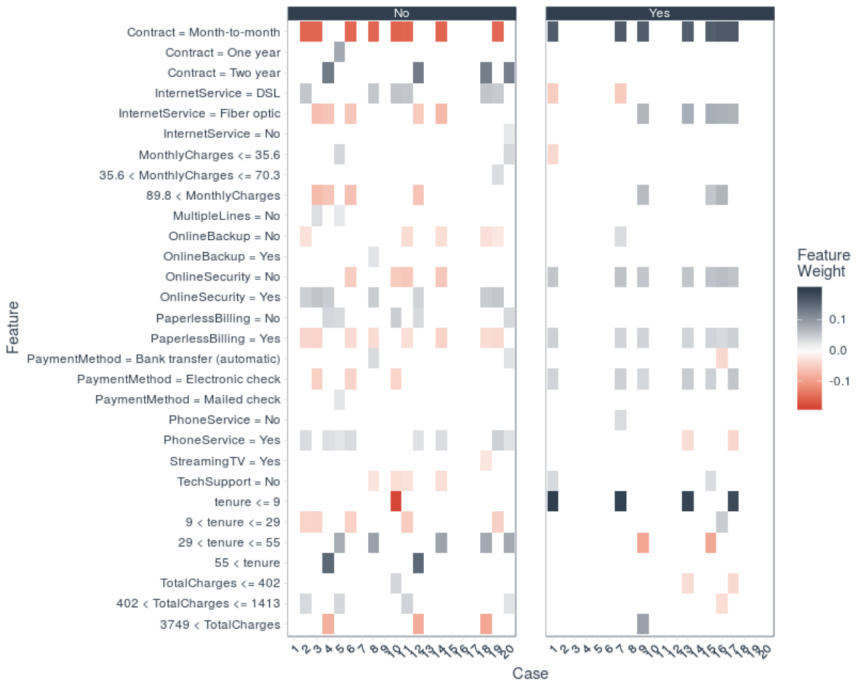

Figure 11: A Heatmap of all variables in respect to churn based on LIME framework.

This heat map is an insightful illustration of what specific variables impact whether a customer churns a not. If the feature weight shows a blue color, is it supporting the variable and the red color shows the vice versa. Clearly, short contract duration does contribute to more probability of churning. Other variables contribute to higher probability to churn are high monthly charges (more than 89$), no security, paperless billing, and low tenure. High total charges which are more than 3749$ also contribute to churn. This is coherent with our hypothesis, in which we hypothesizes tenure, monthly charges are important factors contribute to people churning.

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 12: The cost and revenue break down of each classification rate based on the confusion matrix.

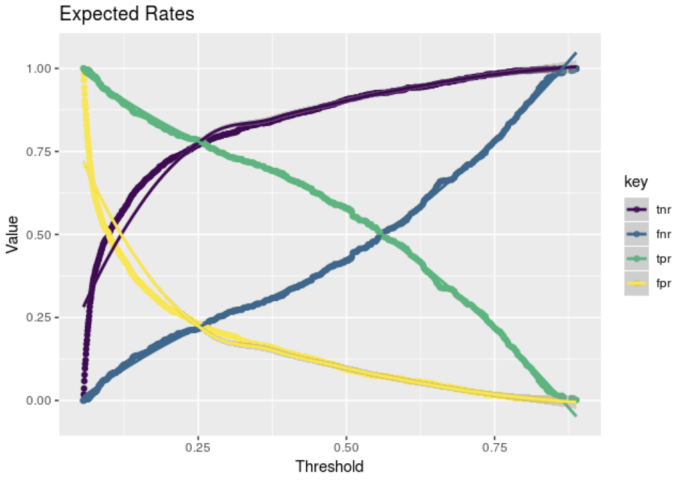

Figure 13: A boat.

8.1 Threshold Optimization

9 Conclusion

9.1 Business Recommendation

9.2 Limitations

References

[1] Ali Tamaddoni Jahromi, Stanislav Stakhovych, and Michael Ewing. Managing b2b customer churn, retention and profitability. 43(7):1258–1268.

[2] Customer value management: The path to profitable growth in telecom.

[3] Emily Jackson. Big telecoms are spending more cash to keep customers, but some tactics raise concerns | financial post.

[4] Charles L. Bonza. Telecommunications infrastructure industry.

[5] Eva Ascarza, Scott A. Neslin, Oded Netzer, Zachery Anderson, Peter S. Fader, Sunil Gupta, Bruce GS Hardie, Aurélie Lemmens, Barak Libai, and David Neal. In pursuit of enhanced customer retention management: Review, key issues, and future directions. 5(1):65–81.

[6] Kamya Eria and Booma Poolan Marikannan. Systematic review of customer churn prediction in the telecom sector. 2(1).

[7] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. 218(1):211–229.