Predicting Churn to Maximize Profits With Predictive Modeling

By Tam Nguyen

Submitted in partial fulfillment of the requirements for the
Degree of
Bachelor of Arts and Sciences
Quest University Canada

and pertaining to the Question
What makes a good decision?

Tamara Trafton, Ph.D.	Tam Nguyen

Acknowledgement

I want to express my sincerest gratitude to my mentor, Tamara Trafton PhD, for helping and assisting me with my Keystone and my study at Quest. Her guidance, feedback and discipline have pushed me to complete my Keystone and my study at Quest.

I am extremely thankful to my family, including my parents, Nguyen Van Dung and Nguyen Thi Ngoc Dung, my sister Nguyen Thi Hong Nhung, my brother-in-law Tran Binh Luc, and my cats, To and Meo. They have provided me great support and caring throughout all these years in college and respected every decision I made, some crazy, some unrealistic.

I owe my great gratitude for my friends, my roommates at Quest and my friends in Vietnam who have provided tremendous emotional support for me throughout all the my university years. They have inspired me, made me who I am today and cultivated the best part of me.

And finally, my special Nguyen Mai Phuong, for her great offering of the essential lessons I can never have in school, and her constructive criticism for the betterment of me. Without her, my undergraduate education would feel incomplete, for there would be no one who could expose to me my deepest and most vulnerable self.

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 (Putz, Herrán, Tortosa, & Reitenspiess, 2012). The traditional source of revenue for telecommunication industry has been callings, but it's changing to texting, image and video processing due to the decreasing cost of internet accessibility. 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 (Jahromi, Stakhovych, & Ewing, 2014).

Though this is a big industry, most telecommunication companies in developed countries are struggling to maintain profitability (Putz et al., 2012). 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 (Jackson, 2017). 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 (Bonza, 2017).

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 (Putz et al., 2012) (Jackson, 2017). On the customer side, it is very easy to switch to a new provider if it provides a more beneficial offer (Canale & 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 (Ascarza et al., 2018). 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 (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). On the other hand, losing customers can lead to (1) negative words of mouth, (2) more opportunity cost because of reduced sales (Eria & Marikannan, 2018). 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$ (Jackson, 2017). 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% (Bonza, 2017).

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 main types of churn: (Lazarov & 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 (Mandák, 2018).

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 (Verbeke et al., 2012). 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 (Eria & Marikannan, 2018).

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 (Verbeke et al., 2012). 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 (Lemmens & Gupta, 2017). 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 (Eria & Marikannan, 2018). 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 (Eria & Marikannan, 2018). 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. [1].

2 Method

The project follows the Cross-industry standard process for data mining framework (CRISP). This framework allows us to build a data science project focusing on business results. The figure below shows the steps of the CRISP framework. The arrow shows how each step is followed by another step. The outer circle shows the constant iteration of the process.

Figure 1: Illustration of the CRISP framework.

As CRISP treats data science work flows as iterative processes, the process can grow and progress because it allows practitioners to come 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

[1] https://github.com/tamdrashtri/data-analysis-final-project

metrics in which we want to know besides aspects of data science so CRISP is preferred for our project. Each step of this framework is specified below:

- 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 Kaggle. The data contains of two main types of variables:

- Customers' 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 their service usage behaviors.

3.2 Formula for the cost of churn

	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 across the US is about 1.9% to 2.1% per month (Hughes, 2019). 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.

In this step, the other thing needs to be considered is the cost of churn in the telecom industry for each customer. The formula to calculate the cost of churn is formulated as followed:

$$y = (\beta - \alpha) * \mu * \mathbb{N}$$

Figure 2: Formula to calculate the cost of churn. The outcome is the total cost of churn. Beta here means the life time value of a customer. Alpha means the number of months customer stays until they churn. Mu is the gross profit telecom companies earn from each customer, and N represents the number of customers.

Specifically, customer lifetime in the telecommunication industry is usually 52 months. Customers usually churn in 19 months. Gross margin in the telecom industry is 47% (CSIMarket, 2018) , from that we can calculate gross profit by multiplying the gross margin with the average monthly charge, which equals 35.1 dollars. Therefore, the total cost of each churner is about 1158.3$. Multiplying this number by the number of churners in this dataset, we have 2.15 million dollars as the total cost of churn with only 1869 churners.

3.3 Estimated cost of churn

We hypothesize two variables that are important contributors to churn, Monthly Charges and Contract and plot the estimated cost of churn by these two variables alone using the formula shown above. The reason we choose these two variables is because they are factors that telecom companies can provide incentives to customers (i.e. give discounts). This makes these two variables more important than other variables. In this specific scenario, since Monthly Charges is a continuous variable, it is best to discretise it into groups for better visualization and categorization. The 'Monthly Charges' variable is binned into 4 ranges. The bin having the lowest number (i.e. bin 1) has observations within the lowest monthly charge. Whereas, the highest bin contain observations with high monthly charges within the range of the dataset.

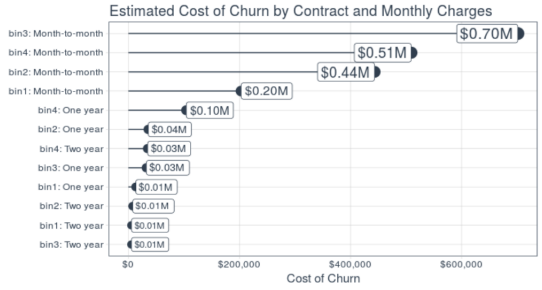

Figure 3: Estimated cost of Types of Contract and Monthly Charges.

The figure shows that Bin 3 with month-to-month contract has the highest cost of 0.7 million dollars since it has the most number of churners, compared to bin 4 having the lower number of churners. We found an interesting pattern that long contract customers often have much lower chances to churn. Therefore the cost of churn by these groups of long contracts is really low. It can be hypothesized that monthly charges and contract are the two very important variables.

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.

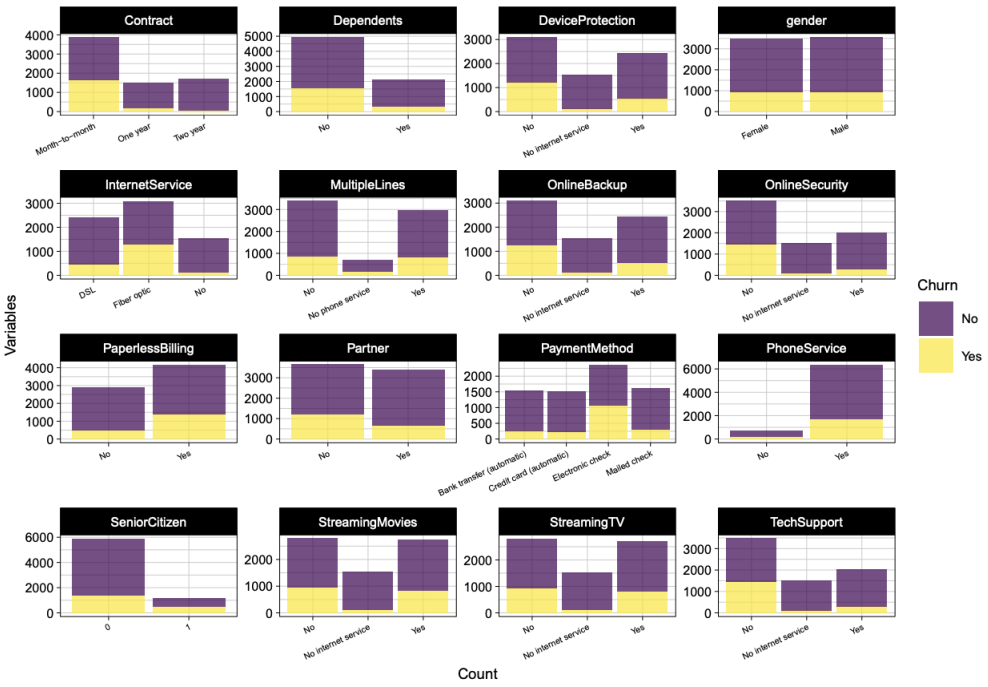

Figure 4: A distribution of numerical variables in respect to churn

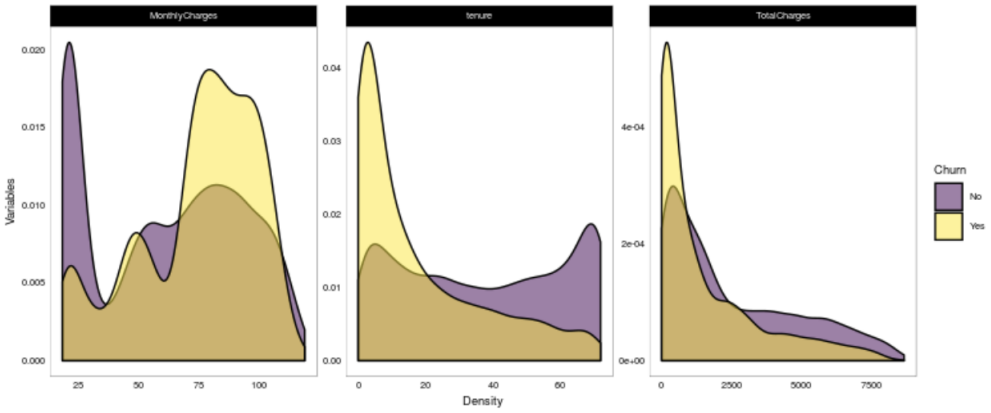

Figure 5: A distribution of categorical variables in respect to churn.

From these two graphs, the first graph shows the distribution of categorical variables and the second graph has the density distribution of the numerical variables. 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 hypothesize 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.

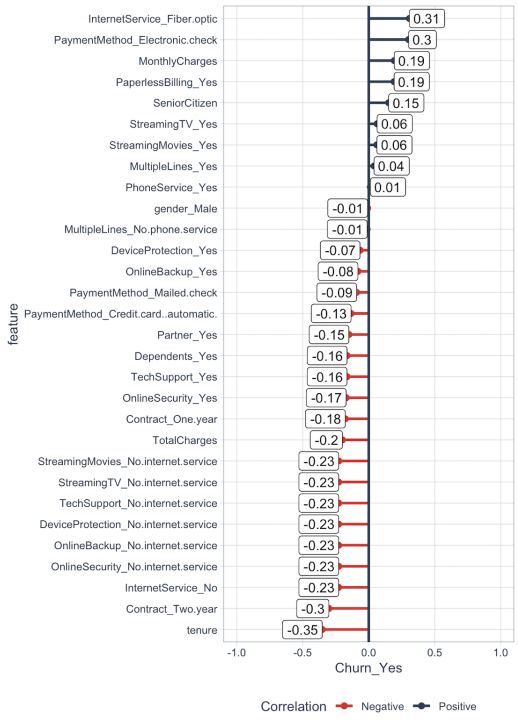

Figure 6: A correlation plot of all variables in the dataset in respect to churn.

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 Modeling

5.1 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 11 observations missing in the Total Charges variable, 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.

5.2 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 implement 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. Cross validation is applied to mitigate the effect of imbalanced data and overfitting problems (e.g. cross validation trains several models on subsets of the input data and evluate them on the other subset of the data) 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.

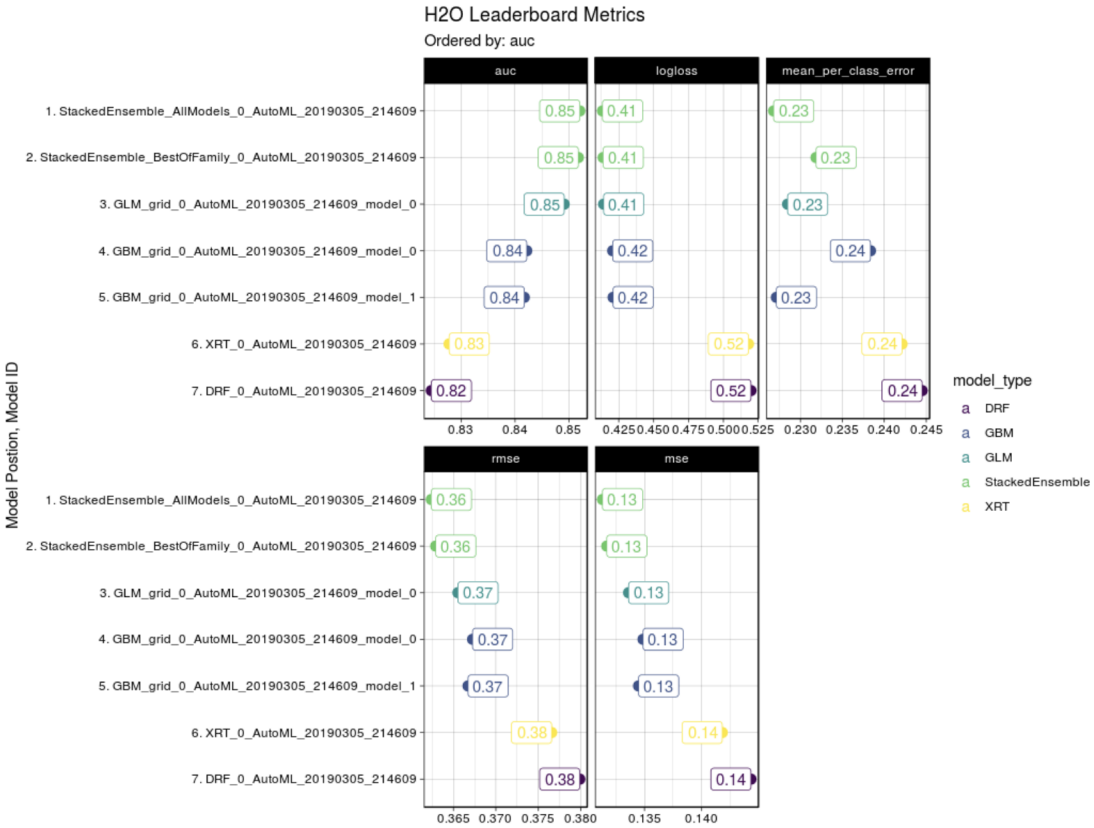

Figure 7: The leaderboard of the models implemented using H2O AutoML ordered by AUC.

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 (*Stacked Ensembles — H2O 3.22.1.6 documentation*, 2018). 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 (Eria & Marikannan, 2018). 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).

6 Evaluation

6.1 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. We see no significant error rates across the confusion matrix. Though the error rate is not low, it is also not very high, so it is acceptable. This confusion matrix will be further used in the later part of the model when we discuss the expected value framework.

6.2 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.

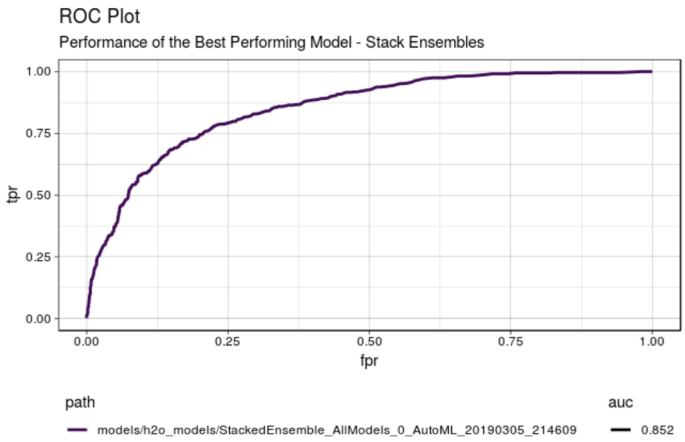

Figure 8: The ROC curve of the best performing model - H2O Stack Ensemble. The horizontal line shows the false positive rate and the vertical line shows the true positive rate.

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.

6.3 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.

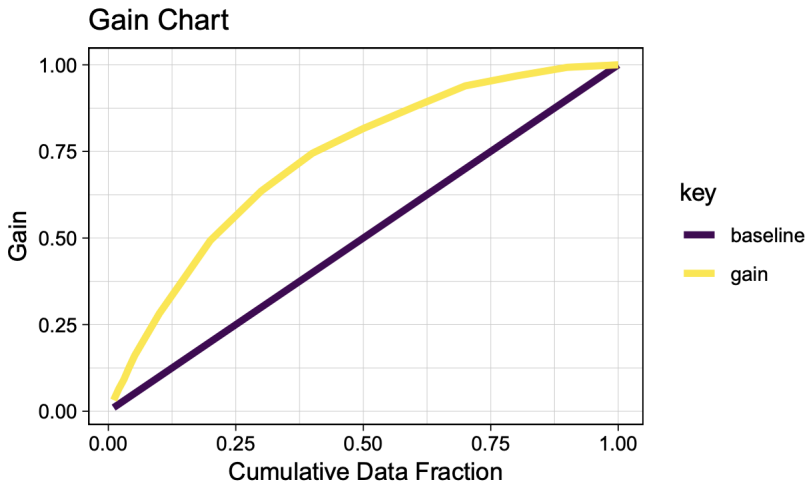

Figure 9: Gain chart of the Stack Ensemble model.

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, Gupta, Kamakura, Lu, & Mason, 2006), thus, life has been used in many churn prediction studies as a performance criterion.

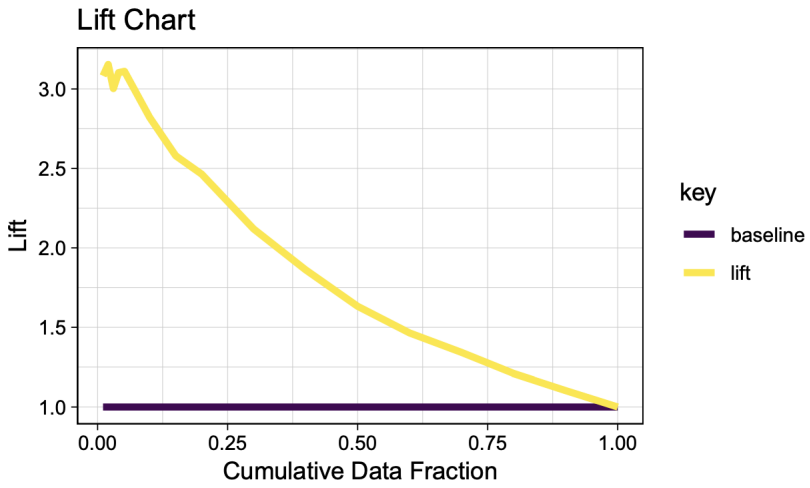

Figure 10: Lift chart of the Stack Ensemble model..

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.

6.4 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 predic-

tions. 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 (Ribeiro, Singh, & Guestrin, 2016).

The general algorithm of how LIME works is (Boehmke, 2018):

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.

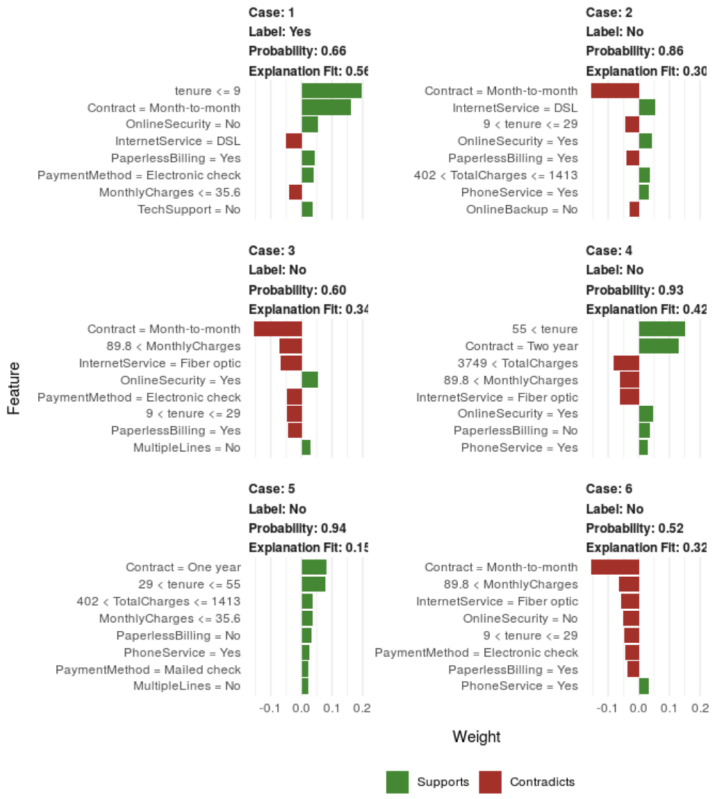

Figure 11: A Multiple Explanation Plot based on LIME framework.

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.

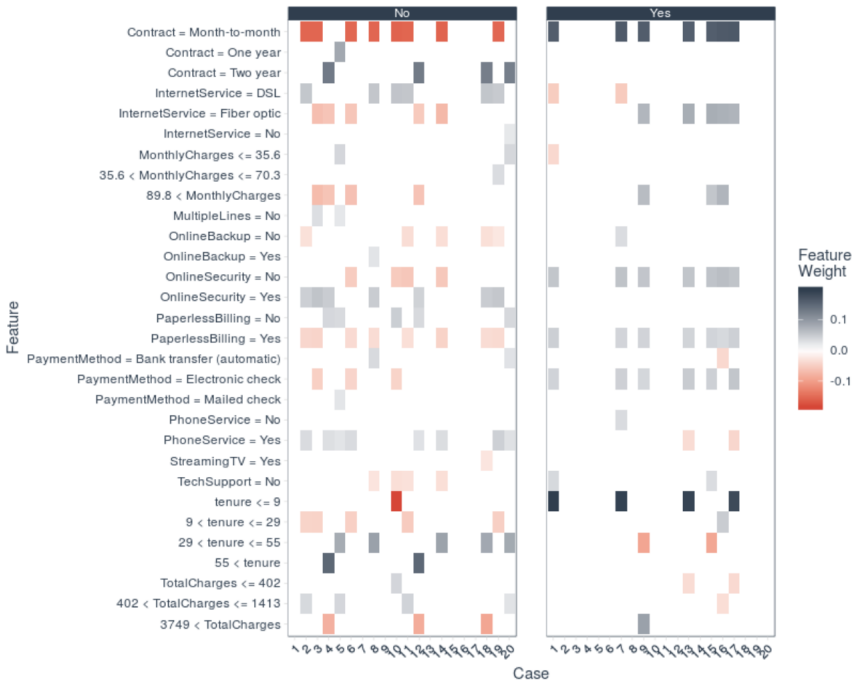

Figure 12: A Heatmap of all variables in respect to churn based on LIME framework.

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.

7 Expected Value Framework

In this section, we will use the expected value framework to calculate the potential profits if implemented the model. The expected value framework is introduced in the book Data Science for Business (2013), aiming at combining the model result with the expected value of the model based on the confusion matrix.

To do this calculation, it is necessary to have the expected rates and the

cost/benefit of each rate, for the formula is as follows:

$$\begin{aligned}\textbf{Expected profit} \;=\;& p(\mathbf{p}) \cdot [\,p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\ & p(\mathbf{n}) \cdot [\,p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot c(\mathbf{Y}, \mathbf{n})]\end{aligned}$$

Figure 13: The expected profit formula (Provost & Fawcett, 2013)

The explanations for each symbol is:

- p(p) is the probability of actual yes in the confusion matrix
- p(n) is the probability of actual no in the confusion matrix
- p(Y, p) is the True Positive Rate (tpr)
- p(N, p) is the False Negative Rate (fnr)
- p(N, n) is the True Negative Rate (tnr)
- p(Y, n) is the False Positive Rate (fpr)
- b(Y, p) is the benefit from true positive
- c(N, p) is the cost from false negative
- b(N, n) is the benefit from true negative
- c(Y, n) is the cost from false positive

In terms of cost/benefit matrix for each expected rates, the table below shows the revenue and cost of each rate:

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

threshold	f1	tnr	fnr	fpr	tpr	b(Y, p)	c(N, p)	b(N, n)	c(Y, n)
0.342	0.655	0.830	0.284	0.170	0.716	1718.72	-2087	0	-368.28

Table 4: The number of each calculation needed for the expected value formula.

The break down of each rate is explained as follows:

- True Positives (Revenue): this is when we give discount to customers we predicted as churn and in response, they no longer churn. We earn 33 months of revenue minus the 15% discount. The total profit is equal 1718.72$
- True Negatives: this is the rate when we predict customers do not churn, so we do not give discounts to them. We earn profits as usual.
- False Positives: this is when we predict they churn and they actually stay. We did give discount to them so we lose profits of the discount. The amount lost is 368.28$.
- False Negatives: this is the rate where we lose 33 months of revenue when we predicted customers stay while they actually churn. 33 months of revenue would be lost since we did not give discount to them, while they can actually change their decision to stay. The total cost of this rate is 2087$.

The F1 score is selected at its maximum score as the baseline and the confusion matrix rates are followed by this. The reason we choose F1 score is because it is used to balance the precision and recall rate. Precision is the total positive predicted observations, while recall is the correctly predicted positive observations to all observations in actual class. They are both useful metrics, but usually when one increases precision, recall will decrease so there is a tradeoff. When F1 score is used, it is to make sure that the best combination of confusion matrix rates are chosen, thus minimizing the classification errors.

Once we have all the numbers for the calculation, using the formula of the expected profit in figure 14, we calculate the profit for each customer and the calculation is shown below:

$$(0.72 * ((0.716 * 1719) + (0.284 * -2087))) + (0.17 * (-368 * 0.17)) = 452\$$$

This shows 452$ as the potential profit per person. Multiply this by the total number of customers in the dataset, we have:

$$452 * 7032 = 3.178.464\$$$

The potential profit by implementing the model alone can generate 3.18 million dollars of the total 7032 persons.

8 Conclusion

8.1 Business Recommendations

The model shows that short contract terms and monthly charges are the significant factors affecting churn. Most importantly, these factors are what is controllable and can be improved by offering incentives to customers. From this example of expected value framework, we introduced one option, which is to give discounts to customers who are predicted to churn based on the confusion matrix. This results in a profit of 3.18 million dollars out of 7032 people in the dataset. Note that each company can have over 1 million people or much more than that, so the profit to mitigate churn when added up can be huge.

Therefore, telecom companies could solve this potential issue of churn by offering special discounts to products or services that are least likely to churn (i.e. phone or data plans that require customers to subscribe with a long term contract with discounts compared to monthly plan or high monthly charges. The higher the charges, the more customers intend to leave and change to a more affordable options. This is time for companies to consider when is best to offer discounts to maximize profits. Personalized recommendations based on customers' profiles are one of the best ways for companies to make sure their churn rate can be mitigated.

Therefore, it is necessary for firms to establish a data infrastructure that has information about customers and their detailed profiles (i.e. professions, age, country of residence, monthly incomes, etc). This can give firms the competitive advantage to develop good predictive models and provide the most suited offer for every customer. The more firms can create a segmentation of customers based on their profiles, the higher the chance of offering personalized products that can mitigate customers' propensity to churn since they would satisfy with firms' offers.

8.2 Limitations

In terms of modeling, there are multiple frameworks and other models can be used. However, in this project, we only choose H2O, so we have few points of comparison with other modeling frameworks. If allowed more time, other frameworks can be applied and we can potentially improve the accuracy of the model by using more advanced feature engineering techniques (e.g. the transformation of explanatory variables to suit to a machine learning model), or by tuning the model's variables using advanced techniques. This paper decides to simplify the

modeling process instead because the dataset chosen is not inherently large and does not have much difficulties in processing.

This project does not look into the technical details of the modeling process. A lot of technical works are handled by H2O framework and therefore it lessens the amount of work needed to be done. Therefore, the mathematical aspect of the model is the limitation of the project since I can only know how it is implemented as an algorithm, but does not know the internal details of the framework. This is a necessary trade-off needed to be made, since this project focuses on the business impact of the model, it is needed to simplify some parts and focus on other parts. However, in terms of trust, we implemented LIME, which allows us to understand how the algorithm predict in terms of individual observations, thereby building more trust into the model.

While other papers usually have multiple churn datasets for cross-checking and see if there are differences in prediction, this project only uses one dataset. This simplifies the process, yet it does not allow us to see how the chosen model can be used to predict other similar data sets. However, there are little telecom churn data that is available publicly, so it is quite difficult to find other datasets. Also, the dataset we chose has already been polished, therefore, there are little pre-processing steps such as cleaning the data needed to be done. Real datasets (i.e. data that is pulled from firms' database) can be much more messy than this. Even if the data is cleaned, it is still quite difficult to achieve AUC score more than 0.8. Thus, the high AUC score of the model in this project is likely to not be similar when applied to real datasets.

References

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., ... Neal, D. (2018). In pursuit of enhanced customer retention management: Review, key issues, and future directions. , *5*(1), 65–81.

Boehmke, B. (2018). *Visualizing ML models with LIME.* Retrieved 2019-03-12, from `https://uc-r.github.io/lime`

Bonza, C. L. (2017). *Telecommunications infrastructure industry.* Salem Press.

Canale, A., & Lunardon, N. (2014). Churn Prediction in Telecommunications Industry. A Study Based on Bagging Classifiers. *Moncalieri, Italy: Collegio Carlo Alberto*.

CSIMarket. (2018). *Wireless Telecom Group Inc (WTT) Gross Profit Margin starting from the forth quarter 2018 to forth quarter 2017, Profitability Trends and Ranking, Fundamental Ratios.*

Eria, K., & Marikannan, B. P. (2018). Systematic review of customer churn prediction in the telecom sector. , *2*(1).

Hughes, A. (2019). *Churn reduction in the telecom industry.* Retrieved 2019-03-18TZ, from `http://www.dbmarketing.com/telecom/churnreduction.html`

Jackson, E. (2017). *Big telecoms are spending more cash to keep customers, but some tactics raise concerns.* Retrieved 2019-03-12, from https://business.financialpost.com/technology/big-telecoms-are-spending-more-cash-to-keep-customers-but-some-tactics-raise-concerns

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing b2b customer churn, retention and profitability. , *43*(7), 1258–1268.

Lazarov, V., & Capota, M. (2007). Churn prediction. *Bus. Anal. Course. TUM Comput. Sci*.

Lemmens, A., & Gupta, S. (2017). Managing churn to maximize profits.

Manďák, J. (2018). Proposal and Implementation of Churn Prediction system for Telecommunications Company.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, *43*(2), 204–211.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.".

Putz, A., Herrán, J. d. l., Tortosa, J. A., & Reitenspiess, M. (2012). *Customer value management: The path to profitable growth in telecom.* Retrieved 2019-03-18TZ, from https://www.strategyand.pwc.com/report/customer-value-management-path-profitable

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Stacked ensembles — H2o 3.22.1.6 documentation. (2018). Retrieved from http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. , *218*(1), 211–229.