Final Project: Predicting Churn Rate of Telecommunication Customers

Team 27: *Ngoc (Stefany) Pham, Niti Doshi, Saim Shujaat, Tiansheng Xu, Ikenna Ugwu*

1. Business Understanding

Our sector of interest is the telecommunication industry, which consists of businesses that facilitate the exchange of information through long distances by electronic means such as telephone, mobile, and wireless internet.

Customer churn, also known as customer attrition, happens when customers stop using the services or products of a business. Customer churn is prevalent in the telecommunication industry because the market is highly saturated and the ability of telecommunication companies to differentiate their products is disappearing[1]. Customer attrition directly cuts into companies' revenue and market share. Therefore, telecommunication companies are relying heavily on analytics to identify customers who have a high probability of churning so that they can introduce appropriate incentives to retain those customers. Retaining customers is usually less costly than acquiring new ones. This is the purpose of our project: to predict which customer would churn in the selected dataset of a telecommunication company.

[1] Jain, Pallav, and Kushan Surana. "Reducing Churn in Telecom through Advanced Analytics." McKinsey & Company. Accessed October 12, 2018. https://www.mckinsey.com/industries/telecommunications/our-insights/reducing-churn-in-telecom-through-advanced-analytics.

2. Data Understanding

Our data is a customer dataset of a telecommunication company retrieved from *Kaggle.com*. The dataset has 100,000 rows with each row representing one customer. There is no duplicate. There are 100 variables including 21 factor variables, 79 numeric or integer variables and no character variable. Our target variable is the binary variable *churn* with value 0 as stay and 1 as churn. Other variables can be categorized into four groups: personal information, usage behaviors, pricing, and phone conditions. There are a lot of missing values for some variables such as *ownrent*, *lor* and *income*. The amount of missing values ranges from 1 to 30,190 values.

3. Data Preparation

To clean the data, we converted categorical variables to factors and recategorized blank values as NAs. Afterward, we proceeded to deal with NAs in three steps:

- First, we run a simple regression with *churn* as the dependent variable and each of the variables with missing values as the independent variable
- If the regression coefficient is insignificant either statistically or economically, we conclude that those variables have little relationship with *churn*. Thus, we drop them from our dataset
- If the regression coefficient is significant economically and statistically, we conclude that the variable has some relationship with *churn*. Those variables are *hnd_price* (handset price) and *avg6qty* (average monthly number of calls over the past 6 months). The amounts of missing values for *hnd_price* and *avg6qty* are 847 and 2839 respectively (less than 3% of the data) Therefore, we delete the rows with missing values of these two variables as deletion of less

than 3% of the data does not affect the power of our analysis significantly. As a result, our clean dataset consists of 93,931 observations and 89 variables.

4. Exploratory Data Analysis

Some other key findings include:

- *rev_Mean* (mean of monthly revenue) ranges from -6.17 to 3842.262. It is interesting that revenue can be negative. Perhaps this could be a result of expensive customer acquisition .
- *months* (number of months in service) ranges from 6 to 61 months. Therefore, we have data of both new and old customers.
- *avgmou* (average number of monthly minutes) correlates negatively with *churn*.
- The usage length variables presented very huge ranges. These variables include those relating to average call length in terms of minutes e.g. *avgmou*, *totmou*, *avg3mou*, *avg6mou*. These huge ranges of over 7000 (in minutes) shed some light on why they had minimal effect on the churn rate.

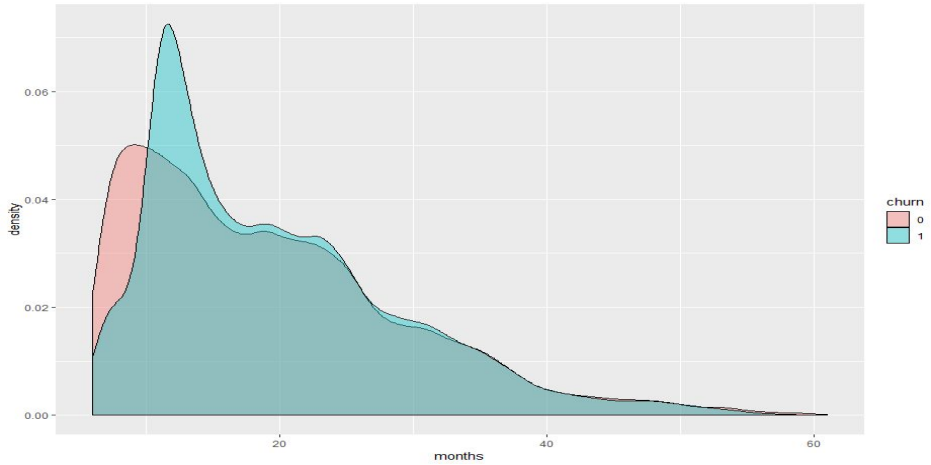

Fig 1: Churn vs Months in service

In Figure 1, we see the relationship between a customer's months in service and churn. During the initial months, churn is quite low and then it hits a sharp peak. This intuitively makes sense as the company might be offering attractive incentives to its customers during their initial months. When those offers expire, customers churn in bulk.

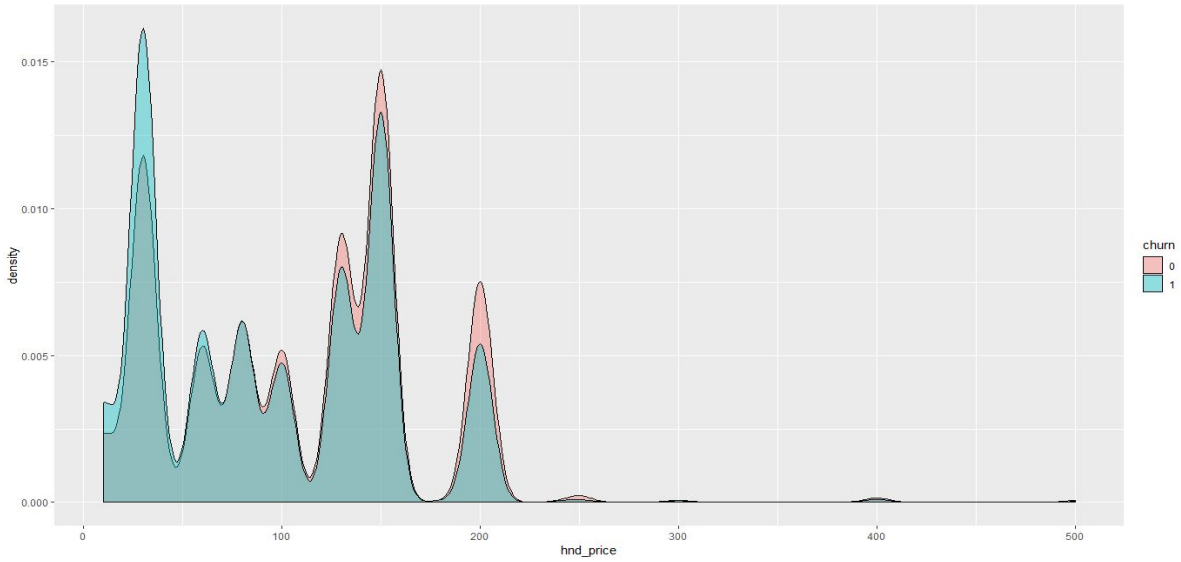

Fig 2: Churn vs Handset price

Figure 2 shows that as price of a customer's handset increases, churn rate decreases. There was some correlation (-0.1) between churn and hnd_price (handset price). A sharp peak in the start indicates customers who churn greatly exceed those who do not, for a cheap handset. After the price of the handset crosses around 80, the trend reverses and the peaks comprise of customers who do not churn. This could be attributed to numerous factors. Perhaps cheaper handsets have poor service/signals that cause the customer to churn from their provider to another in desperate search of better service. Or the fact that customers who can afford more expensive handsets are in general wealthier and are not easily lured into attractive offers by competing providers, hence they do not churn.

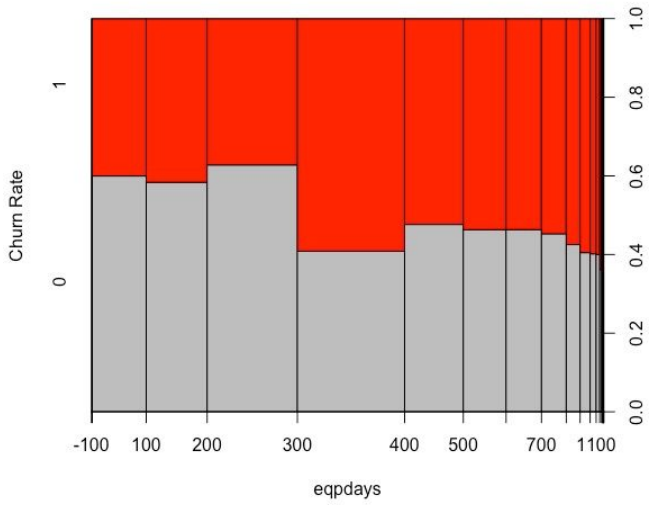

Fig 3: Number of customers churned across total number of equipment days

Figure 3 shows the relationship between churn and the number of days the customer spends with the equipment. The two show a strong relationship and with a mean of 350, we can see that on average, the customers are more likely to churn when the total number of days with the equipment is close to the mean (about 350).

5. Modeling

Our data has 89 variables and therefore, we used Lasso Regression for dimension reduction before we start with the classification. After running lasso regression, we are left with 40 variables, with each categorical variable converted into a dummy variable. We then ran a logistic regression model on the data with *churn* as the dependent variable and all the remaining variables from the clean dataset as the independent variables. Further, we used backward selection to eliminate insignificant variables, along with k-fold cross validation on the training set to get the AIC and model accuracy (in-sample accuracy = 63% & out of sample accuracy = 59.03%). Since, the overall accuracy of the model seemed

low, we tried other classification techniques such as Decision Tree, Random Forest, Neural Networks and Gradient Boosting Machines (GBM). In the end, we chose Gradient Boosting (gbm function in R) as our final model as it gave us the best 'out of sample' accuracy (more information in Evaluation part).

Gradient Boosting: It is a machine learning algorithm that builds an ensemble of shallow and weak successive trees, wherein, each tree learns and improves from the previous one. This proves to be a very powerful technique as these weak successive trees eventually give a strong algorithm, with much improved predictive capabilities.

For our GBM model on the training dataset, we used the gaussian distribution, with learning rate of 0.1, 7000 iterations and max depth of 3. In order to improve the prediction of the model, we also added a cross validation of 5 folds within the GBM method.

Key findings from all of our models:

- Our CART model demonstrates that the only classifier of churn customers is number of days (age) of current equipment (see Exhibit 2, Appendix)
- Based on logistic regression, the longer a customer has been with the company, the lower are the chances of him/her churning
- A customer having a 16-17 year old kid, is less likely to churn
- Unmarried individual is more likely to churn when compared with a married or single individual
- A customer with new cell phone is less likely to churn than a customer having an old cell phone

6. Model Evaluation

The telecommunication company can use our model to identify the customers who are likely to churn and take preventive measures to retain the ones that are on the verge of churning. In the long run, the company can also use some of our insights to target the right customers by identifying the ones that bring the most value to the company.

The accuracy with which our model predicts the customers likely to churn is extremely important as the company will be spending a lot of resources on retaining these customers. Therefore, we have taken the following measures to ensure maximum accuracy possible:

6.1. k-Fold Cross Validation

It is a resampling method used to build a model on the training data in such a way that it can estimate how a model performs on unseen data and therefore, adjusts the model performance to make it robust. This method gives us a model that prevents overfitting on the training dataset and thus will perform relatively better on the test data as well.

For our GBM model, we incorporated cross validation in the gbm() function with 5 folds, meaning that the training data will be split into 5 parts and the model will run 5 times, each time with one part as a holdout and the rest for training. It will then summarize the model using the model evaluation scores.

6.2. Out of Sample Accuracy

For each of our models, we built a Confusion Matrix to get the Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR). Following are the results from each of our model:

	Logistic Regression	CART	Random Forest	Neural Network	Gradient Boost
Accuracy	59.03	57.77	59.09	58.65	62.56
FPR	40.93	53.49	37.97	54.09	62.74
TPR	59.00	68.96	56.17	71.31	62.40

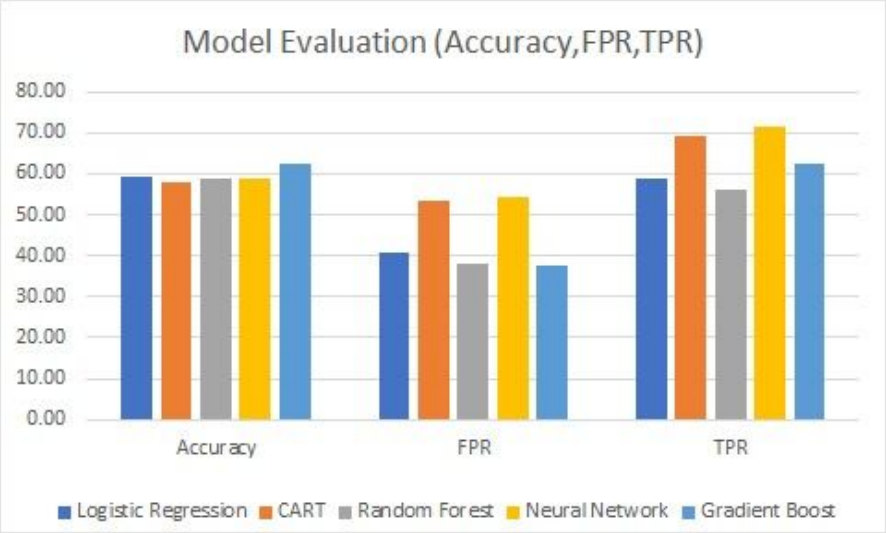

Fig 4: Performance of models by Accuracy, FPR and TPR

As we see, the 'out of sample' accuracy of Gradient Boosting algorithm is relatively the highest and therefore, we choose this as our final model. Additionally, with the lowest False Positive Rate and high True Positive Rate, this model can better predict the customers who are likely to churn. The Area Under Curve (AUC) value of the model, defined by TPR and FPR, is 0.62. An AUC value should ideally be as close to 1, as possible.

Overall, the accuracy and predictive ability of our model are low because the variables that we have, are not as correlated with the dependent variable i.e. *churn*. Therefore, to improve this model, we suggest that the company collect more 'complete' data and relevant variables that correlate more with churn such as competitor monthly charges and promotions flag.

7. Deployment

7.1. Recommendations based on EDA and model findings

From our data exploration, the most impactful variables on customer churn were widely ranged. Some of the most significant variables included *months*, *eqpdays* and *hnd_price.* Our first mode of action would be to tap into the customer's perspective surrounding these factors. For *months*, we have seen that customers churn in bulk after after 10 months approximately. *eqpdays* also reveals that customers are most likely to churn during the mean length of use (approximately 350 days) of their current equipment i.e. cell phone, iPad or other communicating device. Thus, our recommendation would be to target customers who are close to their first year with the company. This can be executed with one year anniversaries reminders, special deals or new perks for customers who are about to turn one.

In line with this, customers are more likely to churn with a cheaper handset. In addition, handset price and length of equipment use are positively correlated, which shows that the customers are more likely to keep on using their current device the more expensive it is. From Figure 3, we see a sharp increase in churn between 250-350 days and a quick reverse from churn to no churn after prices pass 80 in Figure 2. Thus, our recommendation would be to specifically target customers who are about to turn 1 and use relatively cheap handsets (0- 80).

Moreover, based on our model findings, customers who are unmarried churn more. These individuals are likely to be millenials and an important segment of the customer base. Therefore, the company should design packages tailored to millennials' needs in order to prevent them from attriting.

7.2. Cost Benefit Analysis

Some contextual inferences about the Telecommunication industry[2]:

- There are 400M subscribers in the US telecom industry and 92% of adults own a cell phone
- Monthly gross profit per customer is $34
- Monthly loss from churn, per carrier, is 64M
- Acquisition cost for a new customer is $315
- On average, customer lifetime is 52 months and churn takes place at 19 months
- Customer lifetime value is $1,782
- Lost revenue from churned customer is $1,117 (33 months x $34)

Assuming that we offer an incentive to customers we forecast will churn, which is 15% discount beyond the 19thmonth, leading up to the end of their lifetime. We obtain the following matrix:

Cost-Benefit Matrix

	Churn	**Not Churn**
Targeted	($1,285)	$949
Not Targeted	($1,117)	$1,117

The expected value we achieve will be:

*P(Not Churn | X, Targeted)*Value - P(Not Churn | X, Not Targeted)*Value*

From the matrix above we notice that customers who churn despite our offer are the ones causing the greatest financial dent to the firm. Utilizing the accuracy obtained from our model, we will target customers with a strategy that minimizes this segment and ensure not to waste marketing resources

[2] Aditya Kapoor. "Churn in the telecom industry". Accessed October 14, 2018. https://wp.nyu.edu/adityakapoor/2017/02/17/churn-in-the-telecom-industry-identifying-customers-likely-to-churn-and-how-to-retain-them/

on customers our model belies will not churn; because targeting them only causing our revenue to drop.

CONTRIBUTION APPENDIX

	Tiansheng Xu	Niti Doshi	Saim Shujaat	Ikenna Ugwu	Ngoc Pham
Business Understanding	✓	✓	✓	✓	✓
Data Preparation		✓			✓
Data Exploration		✓	✓	✓	✓
Modeling	✓	✓	✓	✓	✓
Evaluation	✓	✓			✓
Deployment	✓	✓	✓	✓	
Presentation		✓	✓	✓	