Customer Churn Prediction on Telecommunication Industry

Yunlei Tang
Purdue University, Daniels School of Business
tang632@purdue.edu

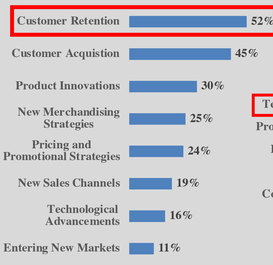
ABSTRACT

This project analyzes the factors that influence customer churn in a telecom company, such as the age of the customer, location of residence, marital status, etc. This is important for a B2C company to predict customer churn because it helps the company to identify and proactively reach out to customers at risk of churn and try to repair the relationship beforehand, reducing the risk of lower revenue for the business. I analyzed which factors have a greater impact on customer churn and brought to the company's attention which types of customers have a high probability of canceling their subscription service. I also built several models such as Lasso, Logistic regression, KNN, and Decision Trees to predict the probability of customer churn. This analysis is important because when the company predicts a high probability of customer churn, it can take action to avoid it.

BUSINESS PROBLEM

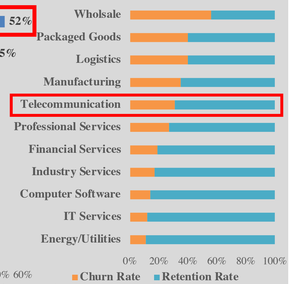
Customer retention has always been the most important factor driving revenue for retail companies. Especially in the telecom industry, customer retention will have a direct impact on a company's overall revenue. Some data show that it costs less to retain an old customer than to acquire a new one, while a 1% reduction in churn can increase profits by 5%. Therefore, the churn rate is a key factor in evaluating the longevity of a company to achieve revenue growth. The main stakeholders of this study are telecommunication companies, and the business goal is to help them accurately identify customers at risk of churn and which factors will impact on customer churn to help them improve customer stickiness.

Most Significant Retailer Revenue Drivers



- Customer retention ranks #1 in retailer revenue drivers, indicating that customer retention is the most important issue that companies should focus on.

Median Customer Churn Rates by Industry, 2022



- Comparing different industries vertically, the customer churn rate in the Telecom industry is at a high position and needs to be addressed urgently.

Source: https://www.revechat.com/blog/reduce-customer-churn/
https://customergauge.com/blog/average-churn-rate-by-industry

ANALYTICS PROBLEM FRAMING

Based on the business problem and representative customer history data, predictive models will be built to explore the relationship between input variables (age, gender, marital status, etc.) and target classification variable (customer status), and figure out which factors have a major impact on customer status to summarize the different types of customer characteristics.

Then, compared with different predictive models to get the most accurate one to predict the customer status of the company in the future, which is Churned, Stayed, based on the customer information obtained.

RESEARCH QUESTIONS

- Which significant factors lead to customer churn in the Telecom industry?
- How well do predictive models help us predict the state of the customer?

DATA

Data Collection & Data Overview

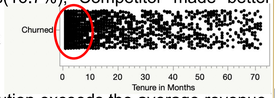
Data is from Kaggle, including 7043 observations and 38 variables, which can be categorized

- Customer information, such as gender, age, marital status, lived city, family status, etc.
- Service purchase, including the type of service purchased, usage, and amount.
- Customer status at the end of the quarter, including reasons if customers are lost.

Although there are so many variables, customer information, broad categories of services selected (Phone/Internet), total revenue, and customer status should be prioritized.

Exploratory Insight

- Top 3 Churn reason: Competitor had better devices(16.7%), Competitor made better offer(16.6%), Attitude of support person(11.8%).
- Tenure vs Churned: Churn rates are higher in the first six months of customer purchase.



- High-value Customers Profile

We define High-value customers as those whose contribution exceeds the average revenue. In this case, we can summarize they are married people (67.7%), are more interested in offer A (19.5%), B (23.9%) or None (52.1%), all more detailed services are more than 50% subscribed, preferring long-term contracts (31.3% for one year, 50.4% for two years).

- Churned Customers Profile

They are unmarried people (64.2%), offer E (22.79%) and None (56.25%) are more popular, and more than half of people would not consider detailed services even though they have purchased broad categories of services, preferring monthly payment orders (88.5%).

Data Preprocessing

- Missing Value & Outliers

Since Telecom company mainly offers two services and customers may not purchase all, there appears null when recording more detailed services. Therefore, I replace the null with "0" if it is a numeric variable and "No" if it is a categorical variable.
I select all the numeric variables, like age and total revenue. No outliers are detected.

- Handle target variable

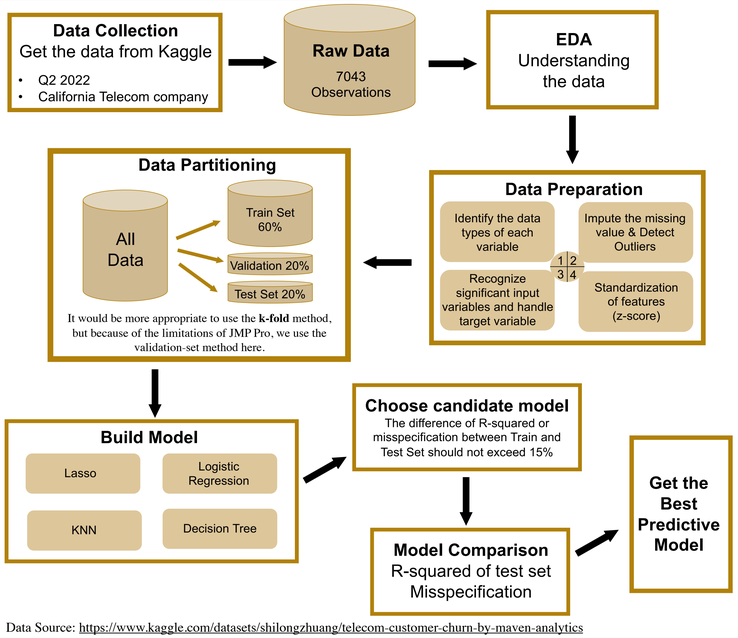
The target variable includes three statuses, churned, stayed, and joined. But the Joined data is useless for our prediction, so I drop Joined and it becomes a binary classification problem.

- Standardization

I standardized all the numeric variables to avoid affecting the effect of machine learning.

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

METHODOLOGY



Data Source: https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics

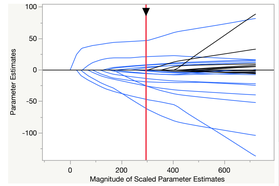
MODEL BUILDING AND EVALUATION – STATISTCIAL PERFORMANCE

Lasso

- Choosing Reason

Because we have 24 inputs, Lasso can shrink unimportant variables to 0 to reduce the variables.

- Statistical Performance
  - Train R-squared: 0.564
  - Test R-squared: 0.531
  - Misclassification rate for test set: 0.1586
  - Significant Variables: Contract (Month-to-month), Tenure in Months(Negative correlation), Number of Referrals (Negative correlation).



Logistic Regression

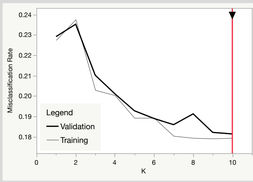
- Choosing Reason

This is a binary classification model and logistic regression is a good baseline model choice.

- Statistical Performance
  - Train R-squared: 0.605 / Test R-squared: 0.565
  - Misclassification rate for test set: 0.1510
  - Significant Variables: Number of Referrals, Contract, Married, Tenure in Months.

KNN

- Choosing Reason: Machine Learning
- Statistical Performance
  - Best k=10 for the lowest misclassification rate: 0.1889
  - Train R-squared: 0.353
  - Test R-squared: 0.314



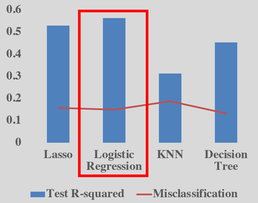
Decision Tree

- Choosing Reason

Easy to understand and explain, can be visually analyzed.

- Statistical Performance
  - Train R-squared: 0.539 / Test R-squared: 0.455
  - Misclassification rate for test set: 0.1335

Statistical performance measurement I use is R-square, which represents the proportion of variance of the dependent variable explained by the variable in the regression model. If the R-square difference between the training set and the test set is greater than 15%, the model is overfitting. For these four models, all are candidate models.



Then I compare the R-squared of the test set and the misspecification rate, a larger R-squared and lower misspecification rate is considered as the best model. In this case, Logistic regression is the best model.

Source	LogWorth	PValue
Number of Referrals	58.199	0.00000
Contract	44.867	0.00000
Married	31.188	0.00000
Tenure in Months	22.622	0.00000
Number of Dependents	17.541	0.00000
Offer	9.885	0.00000

In the Logistic model, the number of referrals has the greatest impact on customer status, followed by contract, married, and so on.

These are important predictors for Telecom company to look out for.

MODEL EVALUATION – BUSINESS IMPLICATIONS

The model can be applied to the telecom industry to help them predict the future status of their customers based on their basic information and purchased services, to better target potential high-value customers. In addition, important factors that influence the results of the model can motivate enterprises to make some adjustments and provide better services to retain customers.

Next step, we can cooperate with consulting firms in the future to provide reports to telecom companies to provide suggestions for their business expansion. In addition, we can extend our predictive model to the market to collect more data for improving accuracy.

CONCLUSIONS

It is necessary to have a good prediction of churn so that telecom companies can better identify potential high-value customers and achieve revenue growth.

- ➢ The number of referrals has the greatest impact on customer status, followed by contract payment form and marital status.
- ➢ Our models predict the future state of consumers with 85% accuracy.

- Limitation
  - K-fold is more appropriate for data partitioning.
  - Adding some transformation variables or other features to increase the proportion of variance that the model explains for the variables.