rgpihlstrom / **Project3**

☆ 0 stars 0 forks

☆ Star | Unwatch ▾

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights

main ▾ …

rgpihlstrom Update README.md … 9 hours ago 178

View code

README.md

Author:Russell Pihlstrom

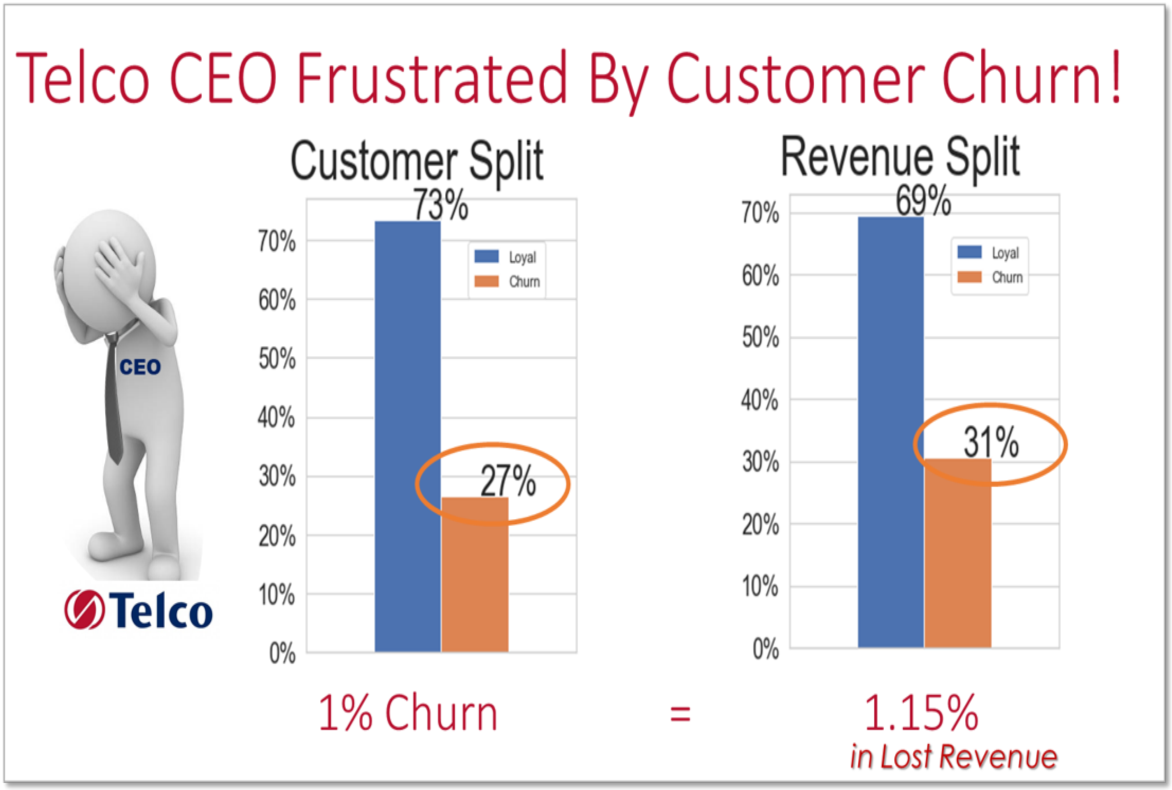
Overview: 85% Churn Detection!

This project uses Decision Tree and RandomForestClassifier supervised learning methods to classify the churn behavior of Telco's customers. By analyzing past actual churn vs. no churn behavior, along with the respective customer attributes associated with each type of behavior, I developed a model that detected 85% of Telco's churning customers. The developed algorithm/"model" can be reused to identify future potential churning customers. The following 4 features were most predictive of churn: **Month to Month Contract, Electronic Billing, Customer Tenure, and Fiber Optics Internet Service.**

Business Problem

Customer retention is a serious concern for all companies. However, within the Telecommunications industry customer churn is of particular importance. Fierce competition, along with difficult to differentiate product offerings makes retaining customers, on factors beyond price, very difficult. Therefore, the historically thin profit margins of the past are only getting thinner, and placing greater than ever importance on customer retention. In fact, a case study done by the Bain Company on the Telecom industry suggested a 5% increase in customer retention can lead to an increase in profits by 25%-95% (https://www.bain.com/client-results/focus-on-customer-engagement-to-improve-retention/).

The Current Situation: Telco is Losing 27% of Customers, 31% of Revenue to Churn.



Business Questions Driving Model Development.

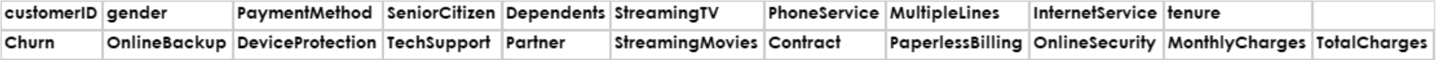
The intended output of this theoretical business case is focused on helping Telco's CEO and Marketing leadership do the following:

- **1. Identify Features Most Associated with Churn.**
- **2. Evaluate Telco's Current offerings Ability to Prevent Churn.**
- **3. Identify Areas for Potential Innovation.**

Furthermore, the model was developed for the purpose of reuse to identify and potentially prevent future customer churn for Telco.

Data

The data used for this project was provided by Telco and published and managed by Kaggle and served as a basis for past Kaggle competitions (Telco Customer Churn https://www.kaggle.com/blastchar/telco-customer-churn) The dataset contains approximately 8k rows and 21 features capturing the purchase, usage, and tenure information on a subset of Telco's customers. 17 of the features are categorical, 3 are continuous, and 1 is ID. A list of features contained within the data are below.



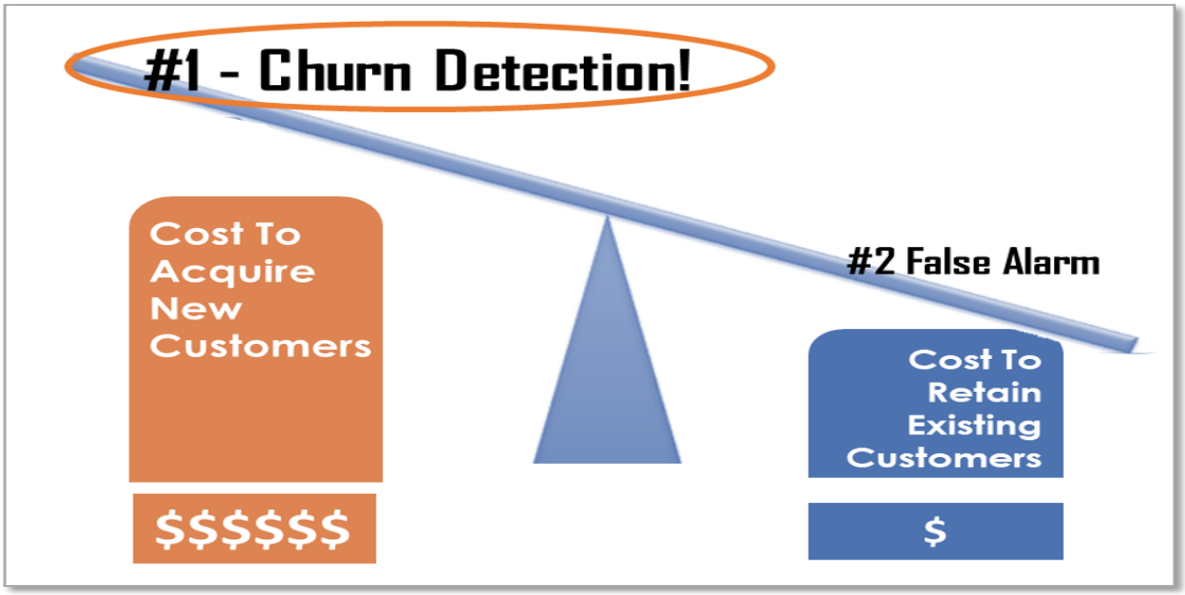
customerID	gender	PaymentMethod	SeniorCitizen	Dependents	StreamingTV	PhoneService	MultipleLines	InternetService	tenure	
Churn	OnlineBackup	DeviceProtection	TechSupport	Partner	StreamingMovies	Contract	PaperlessBilling	OnlineSecurity	MonthlyCharges	TotalCharges

Model Development Methods

This project uses the Crisp DM methodology to generate and optimize the published model. Crisp DM requires blending business strategy, availabled data, and modeling techniques best suited to the business drivers. Model development is and was very iterative. I began by doing secondary research around the basic business drivers of the Telecom industry, gaining a better understanding on the prevalence of churn and the costs associated with fleeing customers. Along with the project requirements noted above, the following additional factors were considered during the modeling process:

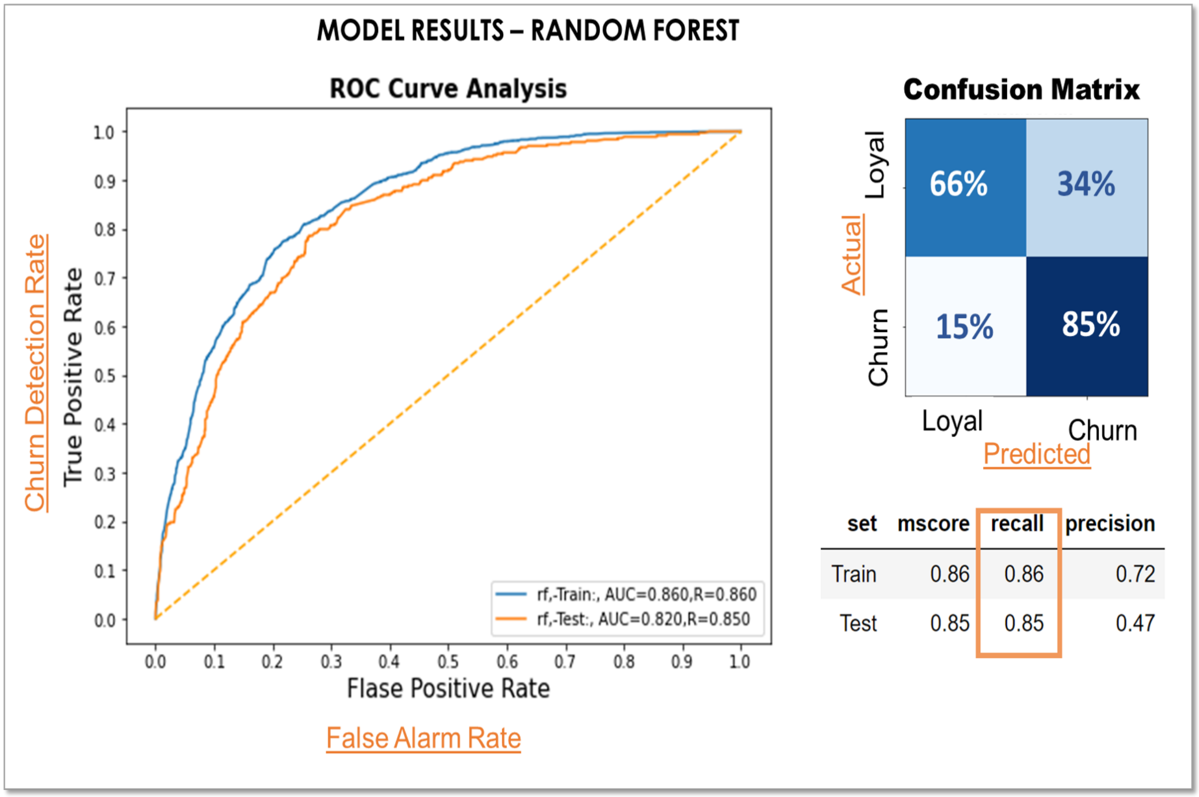
- **1. Data Imbalance** Early in the development process it was obvious that I was dealing with an imbalanced set of data (more information/ rows of data on non-churn customer vs. churn customer). To ensure optimum identification my processes/ modeling would need to account for this imbalance. I addresd this gap by first attempting the model using different weights and ultimately decided to do SMOTE(Synthetic Minority Oversampling Technique) to overcome this challenge.

- **2. Selection of Supervised Learning Classifiers** Initially I tried several different types of classifiers, ranging from Logistic Regression, Naive Bayes, Gradient Boost, Ada, and XGBoost. Ultimately, I decided to use **Knn, Decision Trees and Random Forest** , as these classifiers are non-parametric and are highly interpretable. Interpretability, the disproportionate number of categorical features, along with being able to avoid addressing multicollinearity were the most influential factors in selecting which classifiers to implement for this project.
- **3. Business Drivers: Churn Detection > False Alarms** Recommendations on model development were based on secondary research along with working knowledge on the disparity between the cost to acquire vs the cost to retain customers. In this hypothetical scenario, the CEO of Telco has asked me to place a particular focus on detection at the potential expense of unnecessary outreach activities.


#1 - Churn Detection!
Cost To Acquire New Customers
$$$$$$
#2 False Alarm
Cost To Retain Existing Customers
$

Model Results (85% Detection)

After several iterations, the below recall (85%), accuracy 85%, precision 47%, and AUC82% scores were achieved for the selected classifier.. These results were acheived using the **Random Forest Classifier**. It's important to remember these results were acheived with a focus on recall(detection) over precision (false alarms).

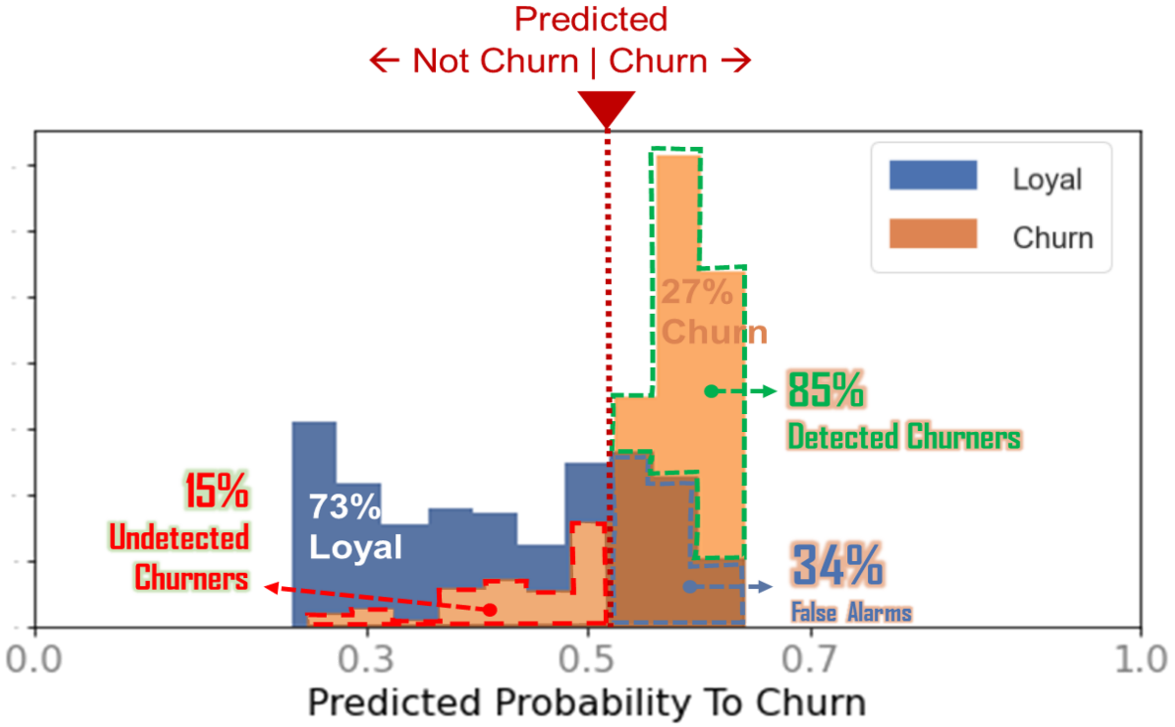


set	mscore	recall	precision
Train	0.86	0.86	0.72
Test	0.85	0.85	0.47

Results Explained (using below illustration):

- **85% Detection** = Model Predicted Churn , Customer Actually Churned
- **34% False Alarms** = Model Predicted Churn , Customer Actually remained Loyal
- **15% Undetected Churn** = Model Predicted Loyal , Customer Actually Churned
- **Red Dashed Line** = This is our THRESHHOLD level (default probability rate) the model uses this % probability to label a customer as churn vs. not churn.

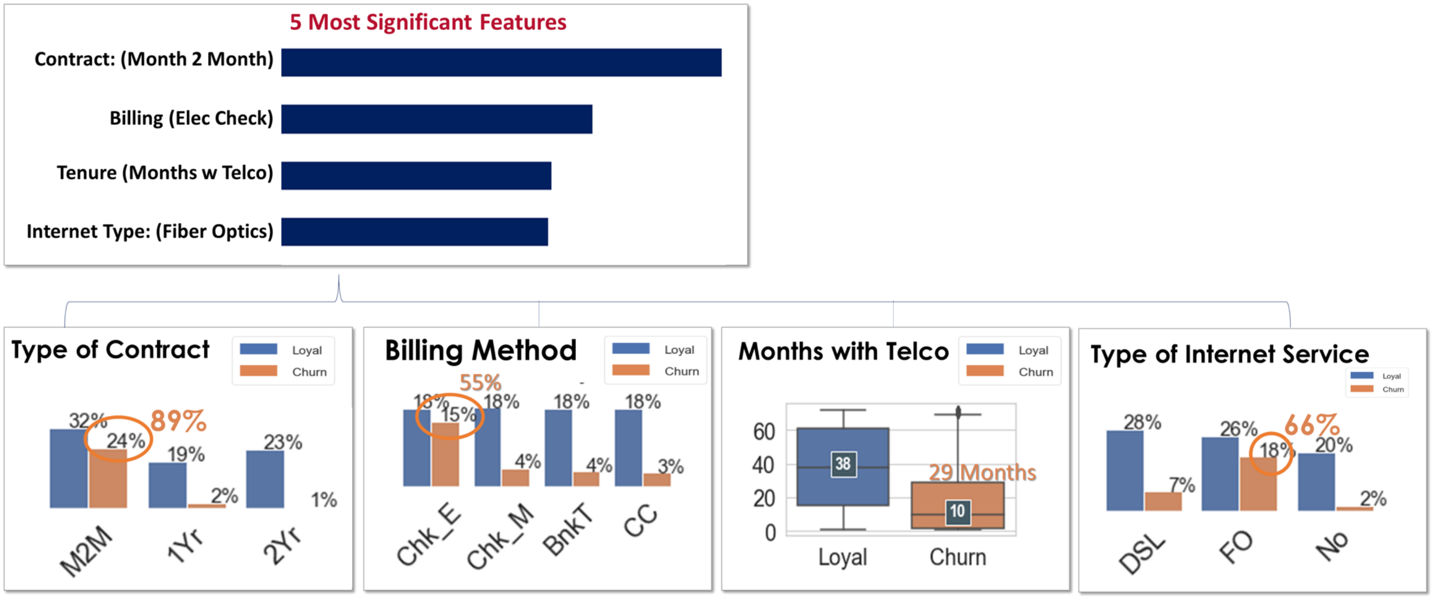
***Key Take Away** = Our model is very good at detecting churn. As we decrease or increase our THRESHHOLD we can capture more or less churners. This in turn results in a higher or lower False Alarm rate. Per my suggestion in next steps below, we need to develop a cost for False Alarms. This will allow us to create a profit tradeoff equation, enabling us to set the "right" threshold for the business.



Business Results/ Recommendations (Using 50% Threshold)

As stated above the goal of the project was three fold. I have outlined and summarized the results of each area of interest below:

1. Top Features Associated with Non Churn vs. Churn:



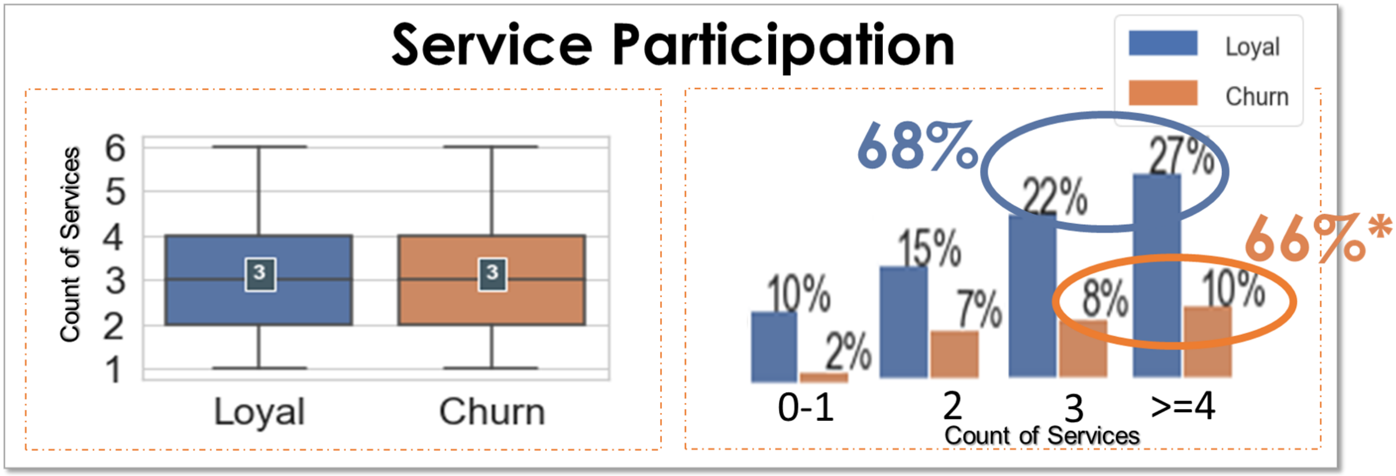
Observations:

- Type of Contract – 89% of churning customers are in Month-to-Month contracts.

- Method of Payment/ Billing – 55% of churning customers pay with electronic check.
- Months with Telco – 75% of churn is occurring within 29 months of becoming a Telco customer.
- Type of Internet Service – 66% of churners are participating in the Fiber Optics Internet Service.

Together these factors were identified by the model as the 4 most predictive of churn. To view a list of all the features used to develop this model see appendix in the pdf stored in this directory.

2. Current Features/ Services Ability to Prevent Churn:

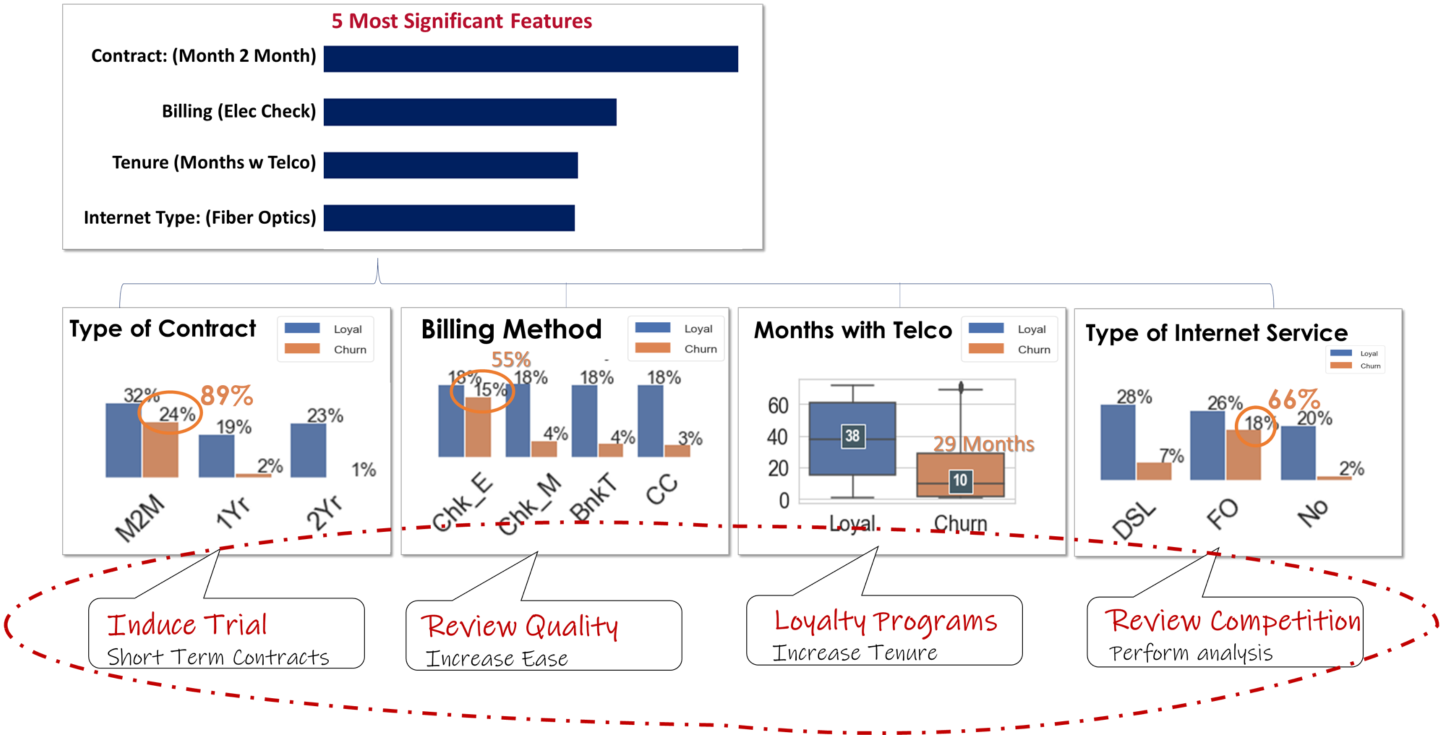


Loyal and Churners Using the Services Equally

Observations:

- Building on above, the feature "Service Count" was not identified as a top predictor of churn. Graphs show equal usage/ participation in services between loyal vs. churn customers. However, ***66% of churners are enrolled in 3 or more services.*** Given that both groups are using services equally and the percent churn is not lower with increased service usage, I would deem that the services are not adequately helping to prevent customer churn.

3. Opportunities for Innovation:

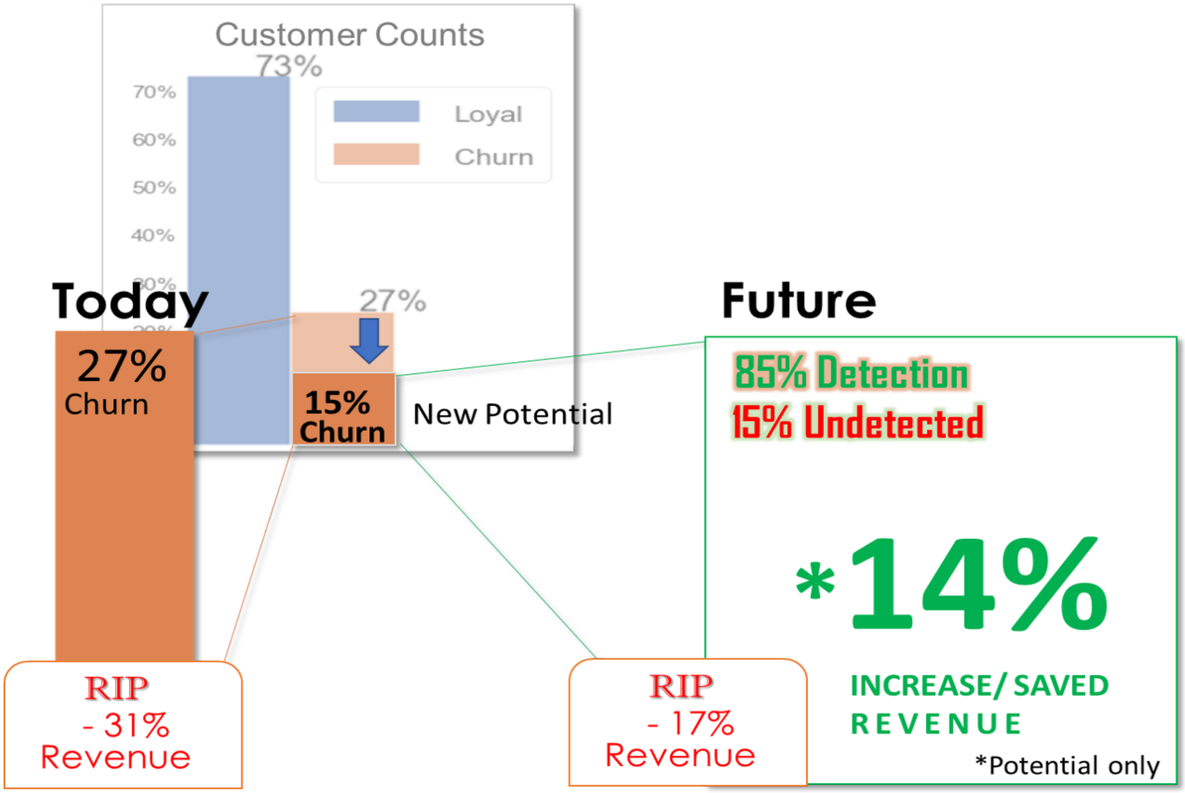


Observations:

My recommendations for innovation are focused on working to create additional solutions around the 4 most predictive features associated with churn which were noted above in point #1.

- Type of Contract – The goal is to reduce spontaneous churn. If churn customers are not interested in a 1 year or a 2 year contract, perhaps there is a shorter term contract that can be created to induce trial. Perhaps try 6 months or quarterly contracts.
- Method of Bill Pay - Examine quality and/ or customer experience required to pay bill electronically. Look to make this process as easy as possible.
- Increasing Months with Telco - Examine ideas for Innovation around loyalty programs to incent longevity
- Type of Internet Service - Given the big difference in rate of churn across DSL vs. Fiber Optics churners, I believe there is something obviously wrong. My recommendation is to engage in a competitive and or quality analysis to ensure the quality of the fiber optics lines are meeting customer expectations.

Potential Impact on Revenue (Assuming 100% Detected Churn Reversal)



Observations:

- Today Telco experiences 27% customer churn rate which = 31% lost revenue.
- Given the **85% churn detection rate** generated by the model, Telco has an opportunity to reduce its churn to as little as 15% (for illustration purposes assuming 100% churn reversal). The result of detecting and reversing the 27% historical rate of churn would result in a savings of 14% revenue.

Next Steps

- **Develop Hard Numbers for the Cost of False Alarms** - In this hypothetical scenario we were not given the cost of falsely reaching out to a loyal customer with a particular outreach/ marketing program. Once definitive numbers can be defined, we can reexamine our Threshold levels.
- **Develop Threshold Evaluation Formular** - Once aligned on costs and savings associated with churning customers, a formula can be created to optimize economics between Detection vs. False Alarm.
- **Examine Detection vs. False Alarm Tradeoffs** - Given the high cost of customer acquisition vs. customer retention some additional analysis may reveal lowering our threshold from 50% to perhaps 40%, which would capture additional undetected churners.
- **Examine Additional Classifiers** - For this project I settled on Random Forest, however, given advances in classifiers such as extreme boost and others, there may be additional opportunities to improve our churn detection rates.
- **Put Model(s) Into Productions** - Once we have optimized our models and/or generated enough models to account for the wide variety of churn data, I would look to automate and deploy the

models via a web-based interface and make it available to the marketing and or customer service teams.

For More Information

See the full analysis in the Jupyter Notebooks or review our Presentation.

For additional info, contact me here: Russell Pihlstrom

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%