Improving TELCO's Customer Service & Retention Rates with Machine Learning

Mandla Senzanje
Minnesota State University, Mankato
120 Alumni Foundation
mandla.senzanje@mnsu.edu

Dr. Rajeev Bukralia
Minnesota State University, Mankato
225 Wissink Hall
rajeev.bukralia@mnsu.edu

CCS Concepts
• **Computing methodologies → Machine learning → Machine learning approaches → Logical and relational learning**.

1. ABSTRACT
The purpose of this research is to analyze an example of how businesses can benefit from applying machine learning principals to their data. Feature importance and finding an effective algorithm to predict customer subscription outcome will be the focus of this research paper.

2. INTRODUCTION
Understanding customer needs has always been a pivotal part of customer care. As competition, demand and technology evolve, it is becoming more important for businesses to evolve.

Businesses are aware that, retaining a customer is easier to accomplish than gaining a new customer or regaining one they once lost [1]. Businesses leveraging the power of Data Science to stay ahead of the competition is essential to survive in today's world.

Initially I will go through a feature selection process to help with the machine learning and predictive model sections. To increase both the accuracy, and training speed of the predictive model, it will be crucial to understand how variables contribute towards the models.

For building my predictive model, I will run multiple machine learning algorithms and get their Confusion/Coincidence Matrix and Area Under the Receiver Operating Characteristic Curve (ROC AUC). This is done to understand which algorithm has the best score.

3. RESEARCH PROBLEM
From TELCO customer dataset, what are the key attributes/features that determine whether a customer will end their subscription or not? (What affects subscription status the most?).

Knowing what affects subscription status, can we create a system that can predict whether a customer is more likely to cancel their subscription or not?

4. METHODOLOGY
4.1 Dataset
The TELCO dataset [2] was obtained from IBM, Watson-Analytics website. The data has 21 attributes, with the target variable being Churn (Whether the customer churned or not (Yes or No)). I picked this dataset because it is a well-known and has been tested with Watson Analytics.

4.2 Exploratory Data Analysis (EDA)
List of some of the libraries used for the EDA, NumPy, pandas, seaborn, and matplotlib. [3] The data cleaning involved minimal work, with exception of one column, 'TotalCharges', which had a few empty values.

The following pair plot was created after I created dummies of the following dataset's columns, ['gender', 'partner', 'dependents', 'phone service', 'paperless billing'].

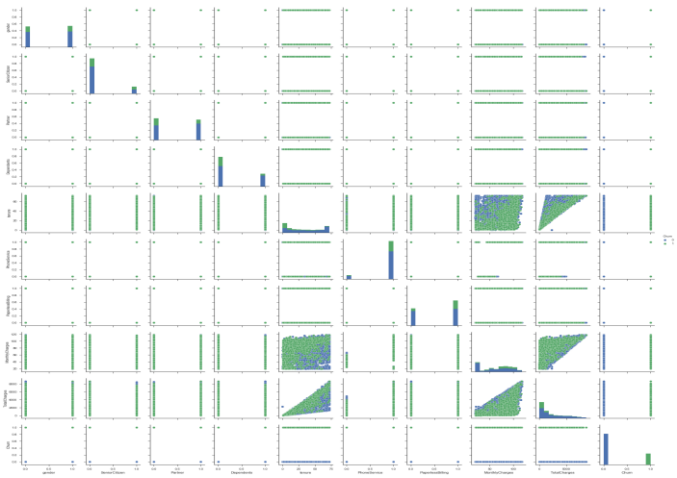

Figure 1: Pair Plot of initial numerical values with Churn as the hue.

4.3 Pearson Correlation
With most of the attributes being categorical values, it was important to convert them into binary form, to make it more appropriate for analytics. After using pandas to create dummies and getting the Pearson Correlation, I wanted to know how well each attribute correlated with weather subscriptions was cancelled or not (Churn).

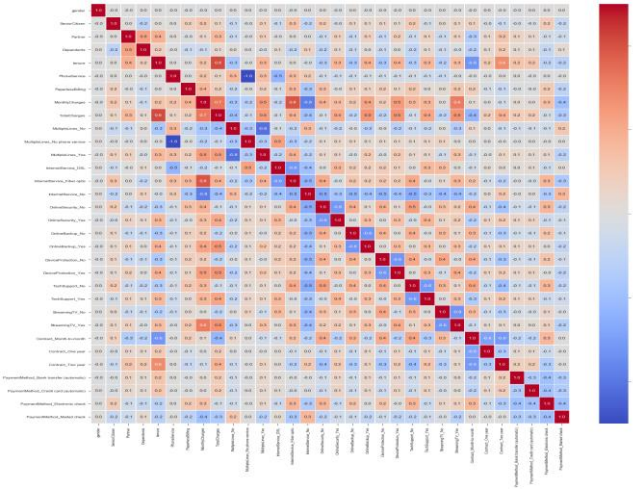

Figure 2: A heatmap of Pearson Correlation, after creating dummies for categorical attributes.

5. MODEL & FEATURE SELECTION

The research problem (What affects subscription status the most?) is a classification problem. Machine Learning tools are used to train the data, which is then used to rank the features [4] and ultimately choose the best model for the prediction system. Algorithms used, Random Forest Classifier, Support Vector Machine, Decision Tree, K Neighbors Classifier, and Gaussian Naïve Bayes. Random Forest Classifier had the best scores, 0.81 precision, and 0.82 recall.

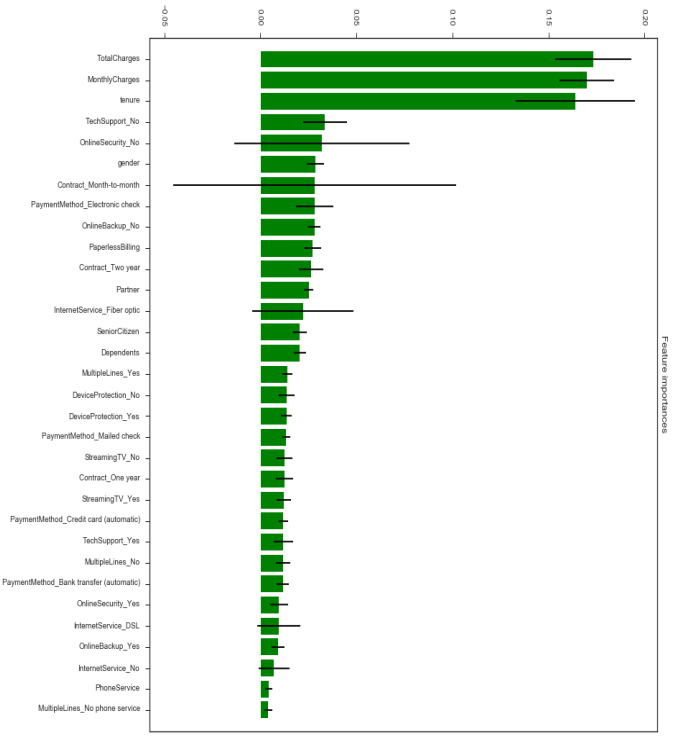

Figure 3: A graph ranking the attributes that affect Churn (feature importance graph).

6. PREDICTION (FUTURE WORK)

With the feature selection complete, the next step is building a system that can predict whether a customer is more like to cancel their subscription of not.

A cross validated score graph was made to see the best number of features to use for our predictions. 32 features came up with the best cross validated score for our predictions [fig 4].

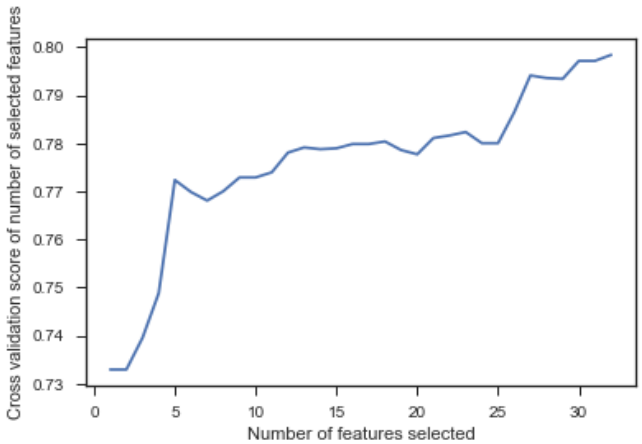

Figure 4: Cross Validation Score graph to see the best number of attributes for our model

7. FINDING & RECCOMENDATIONS

Listed below are a few of the finding from my research that can be applied in decision making for the business:

- Total charges, monthly charges and tenure contributes are the top 3 features when it comes to Churn status. Were we see customers who cancel their subscription are at the early stages of their tenure. My recommendation would be for the business to focus more on satisfying newer customer to increase their tenure
- Month to month contract holders are more likely to end their subscription compared to one- or two-year contract holders. Reasoning, month to month contract holder have less of a commitment to stay as customers. Recommendation, businesses should always try to get newer customers to commit to a long-term contract
- Customer without services such as tech support, online security, online back up are more likely to end their subscription. Enticing customer to apply or utilize these services will decrease the likelihood of them ending their subscription

8. CONCLUSION

The purpose of this research is to understand customers. How can a company stay ahead of their competition? Knowing how customers react, companies can put more resources into finding ways to retain them rather than losing them.

Future work for this project will involve improving the accuracy of my algorithms. My precision and recall score are good, but they are not extremely reliable, more work will be done on my feature engineering. Another aspect I will be adding to my project is adding an Area under the Receiver Operating Characteristic Curve (ROC AUC), this will be done to further understand the accuracy of my algorithms.

Keywords

Data analytics: Data visualizations; Predictive model; Customer retention; Machine learning.

9. REFERENCES

[1] Khalid Saleh. Customer Acquisition Vs.Retention Costs – Statistics And Trends. Retrieved September 20, 2018 from https://www.invespcro.com/blog/customer-acquisition-retention/

[2] Anon. 2015. Guide to Sample Data Sets. (April 2015). Retrieved September 9, 2018 from https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/

[3] Jake VanderPlas. Visualization with Seaborn. Retrieved September 20, 2018 from https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html

[4] Jake VanderPlas. Feature Engineering. Retrieved September 18, 2018 from https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering