TELCO: Customer Churn Modeling

IST 418 Prof. Dunham
Group 2: Nick Santangelo & Peter Perminente

Project overview

By leveraging the extensive customer data pool at Telco Systems, our aim is to forecast consumer behavior and effectively reduce the churn rate. Utilizing sophisticated data analysis techniques coupled with machine learning algorithms, we strive to convert historically churned clients into valuable insights that Telco can utilize to refine and optimize its strategies. This approach is crucial in fostering long-term relationships with the existing client base, as it allows for the identification of pertinent patterns and trends, enabling proactive measures to retain customers and ultimately enhance overall service satisfaction. Through this process, Telco Systems not only anticipates the needs and preferences of its clientele but also solidifies its position in the competitive telecommunications market.

Prediction, inference, and other goals

We intend to conduct a comprehensive analysis to determine the primary factors contributing to customer churn, focusing on creating a decision tree model to uncover the most influential variables in this process. Additionally, we aim to delve into the demographic characteristics of customers, including factors such as gender, age, and number of dependents, to discern any correlations with behavioral patterns.

Furthermore, we will explore the impact of monthly and total charges on customer churn rates, investigating how high rates may influence customer retention. Through rigorous statistical analysis, we aim to develop predictive models that anticipate which customers are most likely to churn in the future based on these variables.

Lastly, as a critical component of our analysis, we aim to develop a sophisticated classification model with a high degree of accuracy, capable of effectively predicting which customers are likely to churn and which are likely to remain loyal. By harnessing advanced machine learning techniques and leveraging the insights gained from our comprehensive data analysis, we seek to create a robust predictive model that provides actionable intelligence for Telco Systems. With this predictive capability, Telco can strategically deploy targeted marketing and promotional strategies tailored to the specific needs and preferences of at-risk customers, thereby increasing the likelihood of retaining them within the company's ecosystem. This proactive approach empowers Telco to preemptively address potential churn scenarios, fostering stronger customer relationships and bolstering customer loyalty. Ultimately, by employing data-driven strategies informed by our predictive model, Telco can optimize its retention efforts and enhance overall customer satisfaction, thereby solidifying its competitive position in the telecommunications market.

Data exploration

The Telco Churn dataset offers a comprehensive glimpse into customer behavior and demographics within the telecommunications industry. Spanning 7043 rows and 21 features, the dataset encapsulates a wide array of information crucial for understanding customer churn dynamics. Columns detailing customer churn status, subscribed services, account particulars, and demographic insights paint a detailed picture of each customer's engagement with Telco Systems. Notably, the dataset provides a holistic view of customer profiles, including family demographic information, albeit without specific age data. Instead, it categorizes customers into binary segments based on whether they are Senior Citizens, offering a simplified but insightful perspective on age-related dynamics within the customer base. This dataset serves as a valuable resource for analyzing churn patterns, identifying influential factors, and ultimately informing strategic decisions aimed at enhancing customer retention and satisfaction within Telco Systems' ecosystem.

Interesting/surprising results

During our data exploration process we performed some basic stat analysis on all of our variables. While most findings were rather tame, some were very indicative on the Telco Customer basis. When investigating the Average Monthly Charges feature (chart 1), we found that most customers opt to pay the minimum of $20 per month. Moreover, a significant spike appears at around $80 per month, indicating a specific segment deeply engaged with the provided services. The data highlights a predominant trend towards the lowest payment tier, with an evident cluster of customers invested at a higher level. Overall, the findings underscores the dual dynamics of widespread minimal payments alongside a concentrated group committed to higher service tiers. When investigating the Tenure feature as it pertains to Customer Churn (chart 2), we found that most customers who churn are newcomers to Telco, averaging a tenure of about 10 months. Conversely, those who remain loyal have a substantially longer average tenure, approximately 40 months. It portrays a clear distinction between the churned and non-churned customer segments based on their length of association with Telco. The data underscores the pivotal role of customer Tenure in predicting churn behavior within the Telco service.

Chart 1

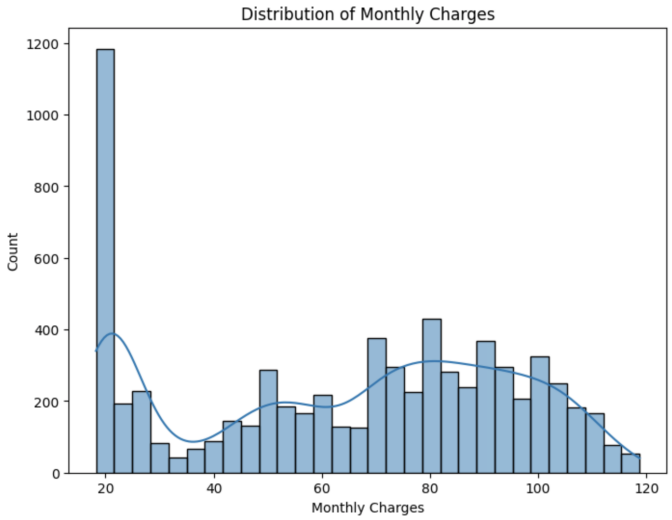
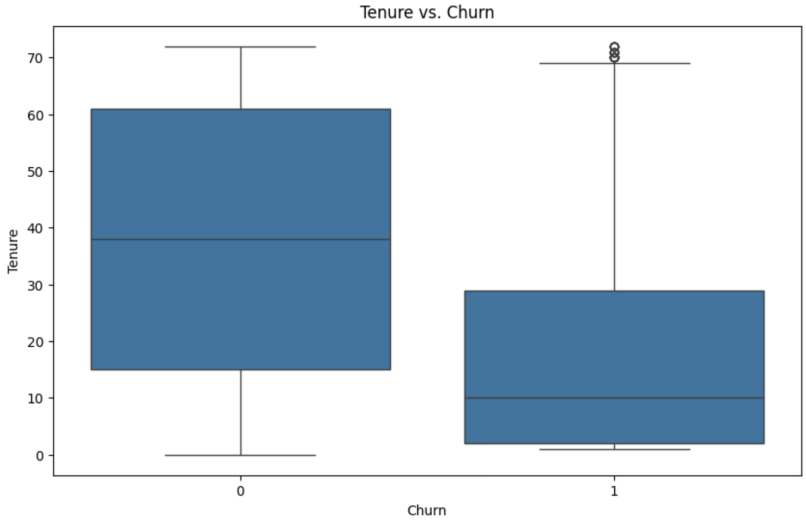


Chart 2



Summary of methods used to solve the problem

When tackling this challenge, we recognized that employing a classification approach would be essential for our machine learning strategy. By employing multiple classification models, we aimed to adopt a comprehensive method to identify the most effective predictor of customer churn within the Telco dataset. Our initial step involved utilizing Decision Trees, providing intuitive interpretability by breaking down decisions into a series of sequential questions. Following this, Logistic Regression emerged as our second method, offering simplicity and straightforward implementation, which was ideal for establishing baseline performance. Additionally, we explored Gradient Boosting Trees (GBT), leveraging ensemble learning to iteratively refine predictions and achieve heightened accuracy. Meanwhile, K-Nearest Neighbors (KNN) relied on proximity-based reasoning, potentially uncovering intricate relationships within the data. Through comparing performance metrics such as accuracy, precision, recall, and F1-score across these diverse methodologies, we aimed to uncover insights into which model best captures the nuanced patterns underlying Telco churn dynamics, thereby guiding strategic decision-making for customer retention efforts.

In addition to employing various classification models, we conducted feature inspection to identify and eliminate less influential features, thereby enhancing model performance. Utilizing a script that inspected feature importance, we used a Random Forest classifier coupled with a VectorAssembler and StringIndexer. By examining the feature importances generated by the model, we ranked the features based on their significance in predicting churn. The systematic process allowed us to prioritize features crucial for predictive accuracy while discarding those of lesser importance, thus refining the model's effectiveness in forecasting Telco customer churn.

Results summary

After testing the plethora of classification models, we found that the clear winner was K-nearest neighbors. The criteria we primarily focused on to determine our best model was Accuracy, False Negatives, and False Positives. These three serve as optimal grading criteria when it comes to providing a comprehensive assessment of predictive performance. Accuracy evaluates overall correctness, while false positives and false negatives highlight specific errors, offering insights into the model's strengths and weaknesses for practical application.

KNN achieved the highest accuracy of 84.50%, with 114 false positives and 197 false negatives, translating to a false negative percentage of 10.54%. Following closely, Decision Tree attained an accuracy of 80.32%, with 175 false positives and 220 false negatives, resulting in an 11.77% false negative rate. Logistic Regression demonstrated an accuracy of 81.42%, accompanied by 130 false positives and 243 false negatives, equating to a false negative percentage of 13.00%. Lastly, GBT achieved an accuracy of 80.52%, with 134 false positives and 257 false negatives, resulting in a false negative rate of 13.75%. These metrics provide a

concise overview of each model's performance, showcasing that KNN is the premier model with the best accuracy, false positive, and false negative rates.

Problems encountered

Some of the problems that we encountered can be related back to our data. We mentioned previously in our data exploration that there was no recorded age in our dataset. Although the models were able to run without age, we felt that age may have been a very important factor in customers churning. Thus if we were to do it again, we may try to either include any external data we may find or possibly try to use imputation methods to fill these missing recorded ages. This may not only improve our results in the future, but could potentially lead to more accurate conclusion explanations for our stakeholder TELCO. Another inquiry we had about our dataset was that there was no mention of users holding a free trial of any of the listed services. Like with Age, a free trial could have been used to understand a lot of early customer churn, as many users who sign up for a free trial may not continue to pay for the services once their trial is up. Thus, this may be something we might want to purpose to TELCO to start implementing in the future and recording data on.

Another problem that we were not satisfied with in our project was with our overall accuracy on our best scoring K-nearest neighbors model. Although our accuracy score wasn't terrible at 84.5%, we wanted to focus on achieving the highest possible accuracy score so we can be more certain in our findings, which can result in better reports for our stakeholders allowing them to make effective business decisions. Thus, in order to achieve better accuracy we could have explored more into improving our current model through hyperparameter tuning, cross validation, or dimensionality reduction.

Summary of how well you achieved your prediction and inference goals

We started this project looking to achieve three goals. The first goal was to analyze factors behind customer churn and identify which variables were most influential. Additionally, we want to explore demographic characteristics to uncover correlations with behavioral patterns. Lastly, we wished to create a sophisticated classification model to accurately predict which customers are likely to churn, enabling Telco to deploy targeted strategies and enhance customer retention.

For our first goal, we found that factors such as Contract, Tenure, InternetService, and TotalCharges have a strong correlation with Chrun. Additionally, factors such as such as Gender, DeviceProtection, StreamingTV, PhoneService, Partner, and Dependents have little-to-no relevance on predicting Telco Churn. For our second goal, we found that there was no identifiable correlation between churn and customer demographic descriptors. Though, the data

set did not have many demographic features potentially due to data privacy laws which prohibit any identifiable demographic info from being publically available.

As far as our last goal, we think that we properly built a classification model good enough to be implemented in everyday business practices. Though, we do wish we had achieved greater accuracy scores and limited false negatives to have less of a real world impact. If we had more time, we think that would achieve these goals and create a much more accurate model. Overall, we believe that we completed all three goals to meet our satisfactions.

Work Cited

BlastChar. "Telco Customer Churn." *Kaggle*, 23 Feb. 2018, www.kaggle.com/datasets/blastchar/telco-customer-churn.