Project 2: Milestone 2 - White Paper
DSC680-T301 Applied Data Science
Joshua Greenert
3/18/2023

Business Problem

The banking and telecommunication industries are notable as having a high rate of churn, which can be costly for companies in terms of lost revenue and acquisition for new customers. The business problem that this project aims to solve is two-fold: depending on the business, there may be unique solutions to each business respectively, and a machine learning model can be created that can accurately predict which customers are most likely to churn based on the industry type. Additionally, this project aims to identify the factors that contribute to the churn for each industry type as well.

Background and History

Customer churn is a significant problem for businesses in the banking and telecommunications industries. According to a study by Bain & Company, the cost of acquiring a new customer can be up to 25 times higher than retaining an existing one (Reichheld, 2014). Additionally, loyal customers are more likely to make repeat purchases and recommend the company to others, which can have a positive impact on the company's bottom line. Previously, companies have used several alternative strategies to reduce churn such as loyalty programs or special promotions. However, these programs were usually focused on a reactive strategy as opposed to a proactive implementation.

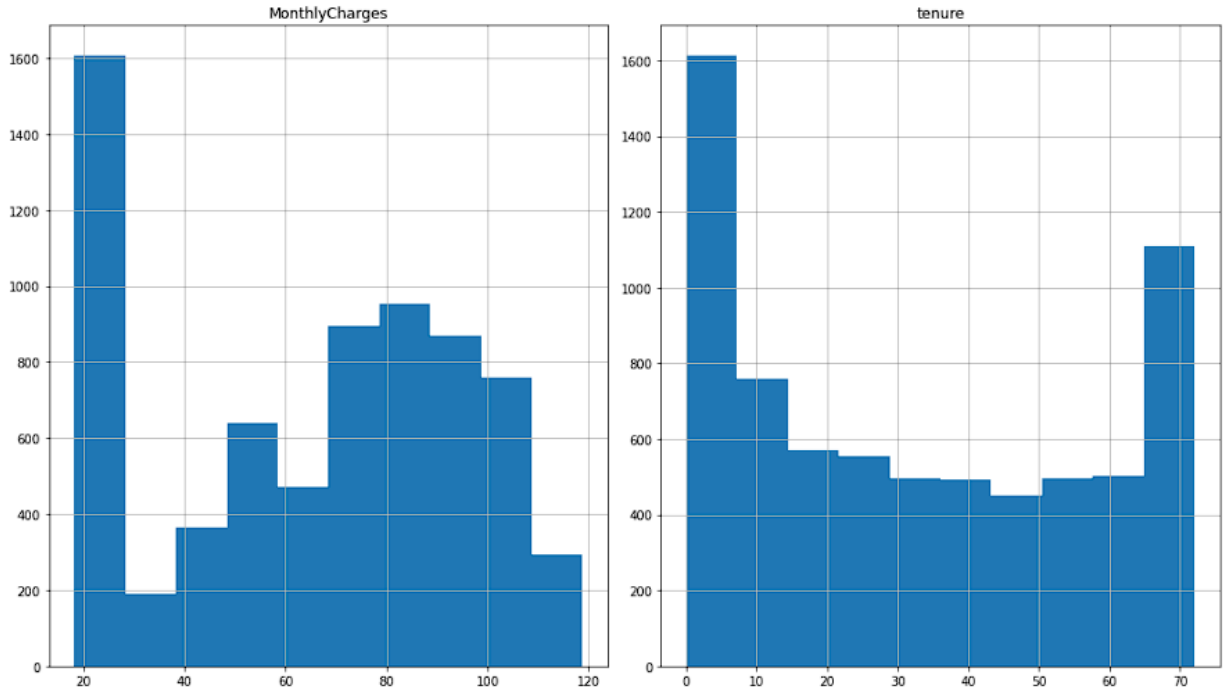
Therefore, companies have begun taking initiative by using predictive modeling to understand which companies are expected to churn before it takes place. Since these factors depend on the industry, two separate datasets for the telecommunication and banking industries will be

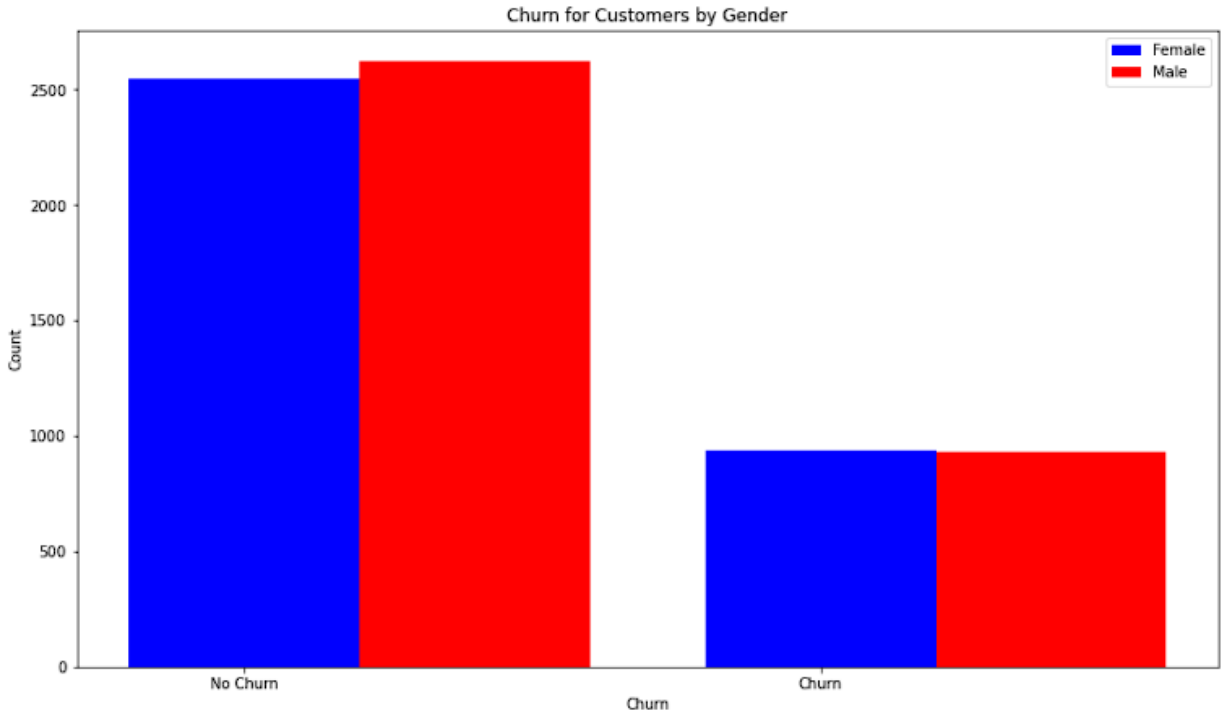
reviewed to identify unique factors that contribute to churn to develop industry-specific solutions to address them.

Data Preparation

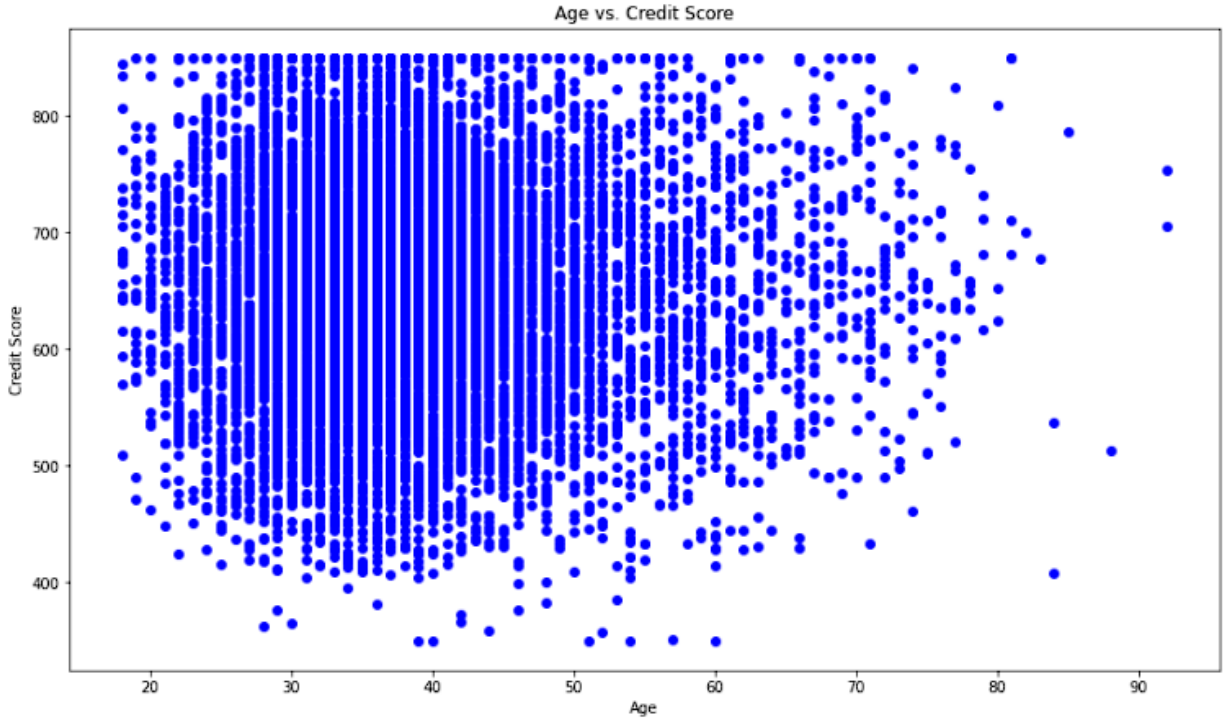
The data preparation process was straightforward for both datasets. The telecommunication dataset required the churn column to be converted to numerical references for “yes” and “no” values; this will ensure that the churn column works as a proper target column for the machine learning modeling that will take place later. Dummies were made for all categorical columns that were intended to be kept, while the columns that held no value were dropped. Similarly, the banking dataset had categorical columns that required dummies to be made as well. The main distinction between the two was the “country” column that was dropped in the banking dataset. This column would have created too many features which could have provided some false positives when checking which features are significant.

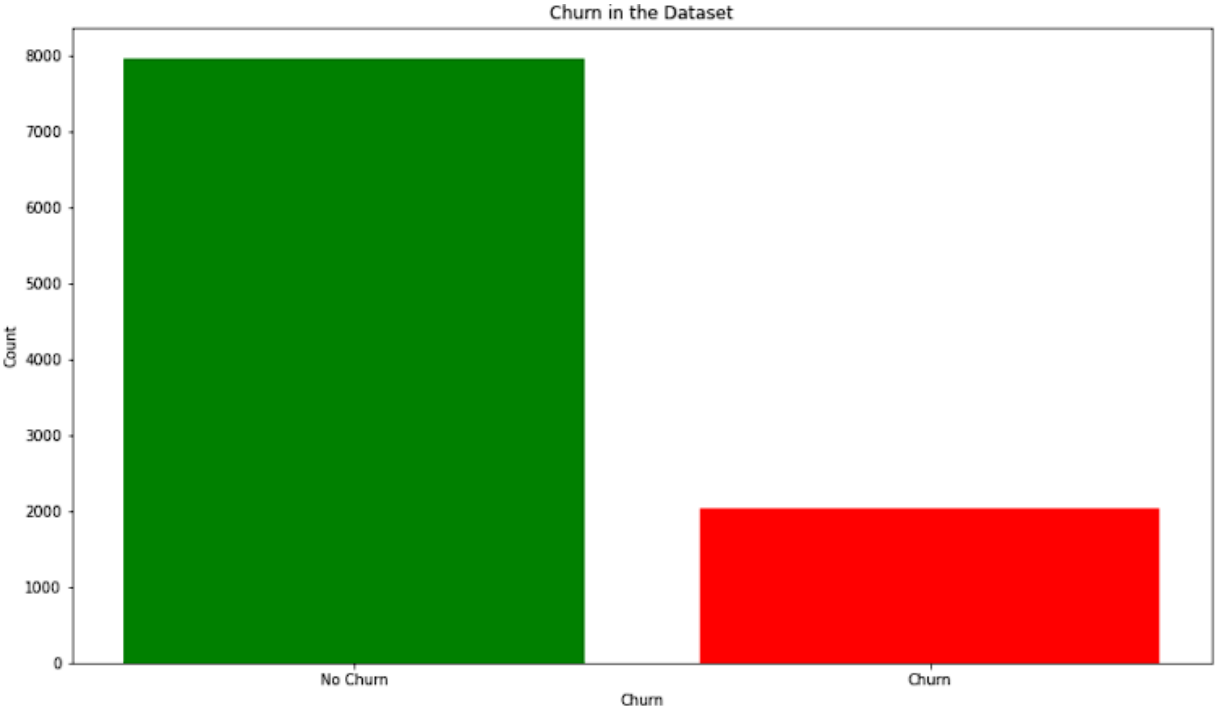
Visualizations





Above are some visualizations from the data provided by the telecommunications dataset. From the monthly charges and tenure, we can see a distributed set of values. Additionally, we can see the churn by gender for customers to discern any notable differences between male or female to ensure that the data will provide a clear direction when using feature engineering to isolate features of interest.





We can review the bank dataset to check for similar data points regarding accuracy and distribution. From the Age vs Credit Score graph, we can see that the center of the data appears near the 35-year-old range with an average credit score of approximately 650. Like the visualizations for the telecommunications dataset, we can see the churned customers and how they compare against the non-churned customers. This data could prove useful in later stages of modeling but helps to confirm uniformity.

Methods

The methods used for this process were logistic regression, decision trees, and random forest regression. While each of these methods may be aligned to the datasets that were chosen, it's likely that the model used for the banking dataset wouldn't be the best choice for the one used on the telecommunications dataset. Therefore, each model was created for each respective dataset, then the metrics were compared against one another to find the model with the best fit.

Analysis

The telecommunications dataset appeared to work best with logistic regression with an accuracy score of 78.96%, recall of 52.14% and AUC-ROC score of 83.23%. All these scores were significantly reduced with the other models created in the same fashion. While these numbers look good on the surface, the recall value implies that the model predicts accurately slightly more than half the time. The main columns that were assumed to have the most potential in predictions were TotalCharges, InternetService_Fiber optic, Contract_Month-to-month, StreamingMovies_Yes, and StreamingTV_Yes.

Alternatively, the banking dataset worked best with the random forest regression model with an accuracy score of 85.85%, recall of 42.75%, and AUC-ROC score of 84.61%. Likewise, this model appears well on the surface but can only accurately predict churning customers slightly less than half the time. The main columns that were assumed to have the most potential with this model are age, estimated_salary, credit_score, balance, and products_number.

Conclusion

Based on the findings in the data, telecommunication companies should review the amounts they have for charges and how those charges are impacting their churn rates. Each one of their services may need to be reviewed against current competitors in the market. It

seems likely that the reason churn is mostly associated with streaming services and contracts could be due to the contract-less nature of more popular streaming services gaining traction in the market. According to a streaming service stat provider FlixPatrol, Netflix, Amazon Prime, and Disney Plus reign as the top three streaming services in the world (FlixPatrol, 2020). To remain competitive, telecommunication companies may need to create agreements with these companies to continue offering products that retain customers between their product lines.

As for businesses in the banking industry, it may be wise to focus more attention on users that are churning within age ranges; from the research conducted, this range appears to be between 29 - 56 years of age. This list of users can be further refined with a focus on the user's estimated salary and credit score to identify potential customers that may require additional attention for retention. Customer service departments could be advised to contact these customers to “check-in” and advise about newer products or ways that users can improve their experience with their current products or services.

Assumptions

The main assumption here is that the data used to analyze churn for banking and telecommunication industries are accurate and representative of the entire population. This may not be the case, and additional work may be required for each respective industry. Additionally, this project has been created under the assumption that the banking and telecommunication industries are different in terms of what model may work best for each industry.

Limitations

The data held within each dataset contains many records, but the number of records could be much higher to provide more insight into churn; moreover, additional data would assist the model in accuracy and its overall predictions. Another limitation is related to time as churn

may change in the future. This current process of analysis may not capture future significant changes, so additional models may need to be tested continuously to analyze customers and their churn potential.

Challenges

The toughest challenge in this project was feature selection since each feature with the highest predicting power would be selected based on the industry. Without additional data, it's hard to determine if these features provide much insight. For example, telecommunication industries have price adjustments that are made consistently with their clients; some companies even offer plans that have a 3-month trial period before the full price becomes active. A detail like this may appear insignificant but could be a large reason for churn as users attempt to service-hop other companies with trail payment plans.

Future Uses and Additional Applications

Either of these analysis methods and models can be applied to other industries that experience high levels of churn, like retail and hospitality. Likewise, the analysis performed can be used by companies to cross-sell or up-sell their existing customers to increase their loyalty while reducing the likelihood of their churn.

Recommendations

One recommendation would be to use this analysis to identify customer segments that are likely to churn while deploying retention strategies. Additionally, companies that intend to implement these strategies should review the cost associated with obtaining new customers in comparison to retaining existing ones. The cost associated with retaining an existing client may outweigh the cost of obtaining newer ones, which can greatly influence the loyalty options provided by most companies.

Implementation Plan

The implementation of this model would require consistent reviews of the working model to include new features as they change within the organization. This model, like any other, is not future proof and would need consistent adjustments to match the values and goals of the business while identifying potentially churning customers. While reviewing retention strategies, companies should also invest more into their customer service departments as these departments, and their ability to satisfy customers adequately, are linked to customer retention and brand loyalty (Mailchimp, 2022).

Ethical Assessment

One potential ethical concern related to the process of predicting customer churn is the use of customer data without the user's explicit consent. To mitigate this ethical consideration, the project will ensure that all data used is publicly available and anonymized. Any findings or models will be shared in a transparent and accessible manner, with clear explanations of the methods used.

Appendix

- **Linear Regression** - A statistical method used to model the relationship between a dependent variable and one or more independent variables.
- **Decision Tree** - A tree-like model used to represent decisions and their possible consequences.
- **Random Forest Regression** - A machine learning method that uses multiple decision trees to make predictions.
- **Accuracy** - A measure of how well a model correctly identifies both positive and negative cases.
- **Precision** - A measure of how well a model correctly identifies positive cases.
- **Recall** - A measure of how well a model correctly identifies all positive cases, including both true positives and false negatives.
- **F1-Score** - A measure of a model's accuracy that takes both precision and recall into account.

- **AUC-ROC Score** - A measure of a model's ability to distinguish between positive and negative cases.

10 Questions from the Audience

1. What are the most effective retention strategies for different customer segments?
2. How do changes in the industry or regulatory environment affect customer churn?
3. What is the potential impact of churn on a company's financial performance?
4. How can companies use data analytics to identify customers who are most likely to churn?
5. What are the ethical considerations when implementing retention strategies?
6. How can companies balance the costs of retaining customers with the costs of acquiring new ones?
7. What are the limitations of using historical data to predict future churn?
8. How can companies measure the effectiveness of their retention strategies?
9. What role do customer preferences and attitudes play in churn?
10. How can companies ensure that their retention strategies follow relevant laws and regulations?

References

BlastChar. (2019). Telco Customer Churn. Kaggle. https://www.kaggle.com/blastchar/telco-customer-churn

FlixPatrol. (2020). Streaming services subscribers., from https://flixpatrol.com/streaming-services/subscribers/

Mailchimp. (2022). The Importance of Customer Service: Benefits and Examples, from https://mailchimp.com/resources/importance-of-customer-service/#:~:text=With%20top%20notch%20customer%20service,free%20advertising%20for%20any%20business.

Reichheld, F. F. (2014). The Value of Keeping the Right Customers. Harvard Business Review. Retrieved, from https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

Topre, G. (2021). Bank Customer Churn Dataset. Kaggle. https://www.kaggle.com/gauravtopre/bank-customer-churn-dataset