**Milestone 4: Finalizing Your Results**

**Customer Churn Prediction for Subscription-Based Services Using Machine Learning**

Author: Jared McLeod

Bellevue University – DSC630: Predictive Analytics

Dr. Andrew Hua

Due Date: July 28, 2024

**Milestone 4**

First, I loaded the dataset so I could manipulate and analyze the data. I handled things such as missing values in 'TotalCharges', some of the items in that column had non-numeric values such as spaces and I ended up replacing them with NaN; I then had to convert 'TotalCharges' to a numeric type. I also converted categorical variables to numerical values, this included things such as gender, partner, dependents, and more. Other things I delt with were dropping unnecessary columns like 'customerID' because it was not useful for this training and I also split the data into training and testing sets, this allowed me to better handle the model's performance. I ended up using the 80 20 evaluations because it is a good balance between training and evaluation.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA close-up of a computer screen

Description automatically generatedA graph with blue and orange squares

Description automatically generated

A screenshot of a text box

Description automatically generated

The model had a remarkable accuracy of 79%, while there is still room for improvement, especially in improving recalling the churn class. Key features include, 'Tenure', 'MonthlyCharges', and 'TotalCharges' are significant predictors of churn and understanding these can help reduce churn. Customers who have shorter tenures and higher monthly charges, are more likely to churn. The decision tree model that was created shows customer churn was predicted with an astounding accuracy of 72%, this indicates that the churn status for 72% of customers was correct, the churn rate was 47%. The relatively low recall for the churn class (47%) suggests that the model misses a significant number of actual churn cases. The really important features when dealing with churn are tenure, MonthlyCharges, and TotalCharges. To better improve my model, I would like to try other machine learning models such 'Random Forest' or 'Logistic Regression'. I would like to monitor the feedback I receive from my classmates to help me better improve this model. I also believe monitoring churn related metrics would also help me improve my model's performance, doing this can help me identify trends and take a more proactive approach and measurements.

**Milestone 3**

**Will I be able to answer the question I want to answer with the data I have?**

The project cultivated is aimed at identifying the key contributing factors to customer churning. The dataset, which includes historical customer data from multiple subscription-based services, encompasses the importances of customer interactions, customer behaviors, demographics, customer feedback, and more. Despite it being a bit of a challenge to find the data, it appears that I will be sufficient in developing a model capable of identifying these factors. The comprehensiveness of the data and the diversity of it all is expected to support the goal of pinpointing the factors which contribute to customer churn effectively.

**Will I be able to answer the question I want to answer with the data I have?**

The visuals that will be useful for explaining the data will be histograms, bar graphs, and most likely scatter plots. Histograms will be used to identify the distribution of customer demographics as well as identify subscription durations. This type of data will better help the audience in understanding customer churn patterns. The bar graph is aimed to visualize the importance of various factors that influence customer churn, such as customer demographic, pricing, competitor influence, and more. Lastly, the scatterplots are here to explore the relationships between variables such as customer satisfaction and churn rates. Doing these aids in identifying key predictors of churn which can be used to help reduce churn rates.

**Do I need to adjust the data and/or driving questions?**

Given the current size and complexity of the dataset, I think implementing the processing steps, such as data cleaning and engineering, can help further my analysis. Exporting different results ensures continuity and helps prevent the need to restart data analysis. This approach will allow for further development while maintaining a focus on the key question which is identifying the contributing factors to customer churn.

**Do I need to adjust my model/evaluation choices?**

The selected models include logistic regression, decision trees, and neural networks, I believe these models are appropriate for analysis. The evaluation metrics being used are accuracy, precision, recall, and F1 scoring; these metrics are suitable in aiding for model performances. However, I have gained new insight, visualizations suggest that adjustments, such as focusing on significant predictors and addressing class imbalances may enhance model accuracy. Dropping irrelevant data points and refining my model parameters may improve my performance models.

**Are my original expectations still reasonable?**

The expectation to predict customer churn and identify contributing factors remains reasonable. The primary challenge potentially lies in managing the complexity of neural network models and processing times. Considering additional data sources and alternative modeling techniques could address these challenges and ensure the project objectives are met.

**Milestone 2**

**Customer Churn Prediction for Subscription-Based Services Using Machine Learning**

**Introduction**

Although companies vary from one another, they all have certain metrics that they try to uphold, they also have data that they collect, one of those pieces of data being customer churn rates. Customer churn is when customers cease to be customers for a certain period of time. Customer churn is a major challenge for companies such as Netflix or Hulu because they are subscription-based services. Predicting when customers are likely to cancel their subscriptions can help companies implement targeted strategies which will make customers want to stay. By adding different retention strategies, companies are able to improve customer satisfaction as well as improve business profitability. The project I am working on aims to help identify the key contributing factors to customer churning.

The success of predictive modeling heavily depends on the quality and the relevance of the data that is being used. For the customer churn prediction project, carful consideration was given when it came to the selection of data sources and the specific data points that were included within the analysis. The primary data source used for this project is historical customer data from a plethora of subscription-based services. The data encompasses various aspects, such as customer interactions and behaviors. The data is collected from a multitude of platforms, by doing this, I have ensured a diverse and a more comprehensive dataset.

**Model Selection**

The different models I plan to use would be logistic regression, decision trees, and possibly neural networks. Logistic regression is relatively simple and helps to interpret models sustainable for classification tasks like predicting churn, it will answer by stating yes or no. Decision trees will provide clear insight into which factors lead to customer churn by visualizing the decision-making process, and neural networks are capable of capturing complex patterns in data. This is beneficial because it makes the data suitable for high-dimensional relationships. The effectiveness of a model will depend on how effective it is. The metric I will be using to measure effectiveness will be accuracy, precision, and F1 score. Accuracy is the proportion of correctly predicted instances out of the total instance count. Precision, which is also accompanied by recall measures the correctness of positive predictions as well as measures the ability to identify all positive instances. F1 score is the harmonic mean of precision and recall, this provides a balance between the two. There are certain things I will be looking at to have a nonbiased dataset, such as, customer demographics, subscription details, and customer feedback. Looking at these details ensures there is an understanding of customer base line, I understand the insight into customer loyalty and payment behaviors, and I will be able to gauge customer satisfaction and identify areas for improvement.

**Learning Outcomes**

The objective and learning outcomes for this project are to predict customer churn rates and develop a reliable model to predict when customers are more likely to churn, as well as understand the main factors contributing to customer churn. Another area I would like to focus on will be providing recommendations and offering actionable insights for companies to improve customer retention strategies. Some of the risks and ethical concerns will be ensuring the confidentiality of customers and their data. Ensuring the data is secure is top priority. I want to ensure that any sensitive data is secure as well. There will also be an inherent bias in the data, whether it is related to gender, age, sex, or any other characteristic, these biases will be identified and addressed accordingly. Lastly, I want to ensure that complex models like neural networks can be interpreted as they can be challenging to understand. I have to ensure that the model is accurate while also providing a clear understanding of the information gathered.

**Risks and Contingency Plan**

If the initial project plan encounters some significant obstacles that are a bit too difficult within the time period we are allowed, ten weeks, I plan to use additional data sources. Integrating additional data sources will enrich the data such as customer feedback surveys. I will also try and figure out ways to shift my focus. If predicting churn proves to be too much of a challenge, the project may pivot to understanding customer satisfaction or identifying high-value customers as a secondary goal.

**Summarizing Project**

This proposal outlines the initial plan for predicting customer churn using a comprehensive dataset and various predictive models. This project aims to provide valuable insight to subscription-based service companies to help mitigate and reduce customer churn. By providing this information, companies such as Netflix or Hulu are able to enhance retention strategies. While the proposal is a starting point, flexibility is maintained to adapt to new findings and challenges as the project progresses.

**References**

Böringer, Jochen, et al. “Insights to Impact: Creating and Sustaining Data-Driven Commercial Growth.” McKinsey & Company, McKinsey & Company, 18 Jan. 2022, [www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/insights-to-impact-creating-and-sustaining-data-driven-commercial-growth](http://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/insights-to-impact-creating-and-sustaining-data-driven-commercial-growth).

“Customer Churn Prediction for Subscription Businesses Using.” AltexSoft, AltexSoft, 27 Mar. 2019, www.altexsoft.com/blog/customer-churn-prediction-for-subscription-businesses-using-machine-learning-main-approaches-and-models/.