**Project 3: Customer Churn Analysis**

Christina Vosnak

Data Science, Bellevue University

DSC 680

Professor Iranitalab

June 1st, 2025

**Business Problem**

Customer churn, or the loss of clients or subscribers, is a critical issue within any industry and business. To maintain profitability, it is necessary, especially in competitive markets, to build and maintain customer loyalty not only to drive financial results but also to build a positive reputation in the market. This project specifically looks at customer churn data from a telecommunication company to develop predictive models to uncover patterns and factors contributing to customer attrition that can be standardized across industries.

**Background/History**

Customer churn has long been a significant concern in the telecommunications industry due to the high costs associated with acquiring new customers compared to retaining existing ones. On average, acquiring a new customer can cost five times more than retaining a current one, which underscores the financial impact of churn (Kotler & Keller, 2016). The telecom market is particularly vulnerable because of minimal switching costs and aggressive competitor promotions, making it easy for dissatisfied customers to leave (Ahn et al., 2006).

Churn prediction models have become more critical with the growth of data availability and advances in machine learning, allowing companies to identify patterns in customer behavior and take preemptive action. Predictive analytics helps organizations segment customers, forecast churn risk, and deliver personalized retention strategies (Verbeke et al., 2012). By incorporating historical usage data, service subscriptions, billing information, and demographic attributes, telecom companies can build a more holistic picture of churn drivers and intervene before a customer decides to leave.

**Data Explanation (Data Prep/Data Dictionary/etc)**

For this analysis, I’ve sourced a dataset from Kaggle which has been provided by IBM to build predictive models and communicate any relevant results. The dataset has over 7000 entries across 21 features related to demographic, account and service information. The final column, labeled ‘churn’ identifies whether or not the customer churned based on the factors provided. As can be seen from the chart below, the data is slightly imbalanced with most customers not churning. This will be addressed at the model stage whether or not to apply under or oversampling.

A blue and orange square with numbers

AI-generated content may be incorrect.

To prepare this dataset, I checked for missing and duplicate values which I found were none so I did not have to address these two common challenges. I also turned all categorical variables into dummy variables to be ready for modeling purposes. I then created a distribution chart for all numerical variables within the dataset to understand the information. Below is the distribution chart for Monthly charges. As we can see, most charges are on the lower amount of $20 meaning that this telecommunication company is distributing services at a low cost overall. However, it is important to note that charges can range to $120.

A graph of a distribution of monthly charges

AI-generated content may be incorrect.Finally, I created a correlation heatmap to see if any linear relationships exist within the data prior to modeling to begin to hypothesize about the information. The most important factor to note, is that there are not many numerical features in this dataset so preliminary analysis of the features is limited until model building.

A screenshot of a graph

AI-generated content may be incorrect.

**Methods**

For this analysis, I will build 3 predictive models to then analyze results and determine a model best fit to predict whether a customer will churn based on the data. I’ve selected a logistic regression model, a random forest classification model, and XGBoost as the three models I will train and test with the information. I will use evaluation metrics to determine the accuracy of the model with precision, recall, F1-Score and ROC AUC to determine whether this model can accurately make real world predictions.

**Analysis**

Exploratory data analysis revealed that customers using electronic check as their payment method, having a two-year contract, and opting for paperless billing showed a higher likelihood of churn. After encoding categorical variables and addressing class imbalance using SMOTE, three classification models were developed and evaluated: Logistic Regression, Random Forest, and XGBoost. All models were trained on oversampled training data and tested on the original test set. XGBoost yielded the best performance overall, achieving a recall of 0.8471, precision of 0.5550, and a ROC AUC score of 0.8291. Logistic Regression performed similarly in recall but had a slightly lower AUC. Random Forest delivered comparable results but did not outperform the other two models. Feature importance analysis from the XGBoost model showed that contract type, tenure, and monthly charges were the top three predictors of churn, suggesting that longer-term contracts and consistent service usage contribute significantly to customer retention.

**Conclusion**

To conclude, all three of these models, didn’t yield high results, with the most highly predictive model being the XGBoost Model. By looking at the ROC Curves, they all are pretty similar with the results of this analysis.

A graph of a logistic and logistic rate

AI-generated content may be incorrect.

To further understand the data, additional classifier models should be tested to develop a more accurate representation of the information. Additionally, to further understand the driving factors for customer churn, top features should be identified to see which ones influenced the model and how significant their prevalence is. For these models, I took the top 3 features of the best performing model which was XGBoost and charted them below. We can see the features with the most importance are customers using electronic check as their payment method, having a two-year contract, and opting for paperless billing however, the importance of these features is very low meaning that there isn’t one significant factor causing churn.

A graph with text on it

AI-generated content may be incorrect.

**Assumptions**

This analysis assumes that customer behavior patterns remain relatively stable over time and that no significant external disruptions occurred during the period in which the data was collected. It also presumes that the dataset is representative of the company’s broader customer base and that all relevant features for predicting churn have been captured in the available data.

**Limitations**

The dataset used in this study presents several limitations. It represents a single snapshot in time and does not incorporate temporal trends, making it difficult to observe customer behavior dynamics over longer periods. Additionally, it lacks behavioral or sentiment-based features such as customer satisfaction scores, support interactions, or feedback data, which could offer deeper insight into churn risk. Overall, the dataset while useful for exploratory purposes, is not an accurate representation of churn analysis over time across a full customer base.

**Challenges**

Key challenges in the analysis included addressing the significant class imbalance between churned and non-churned customers, which required the use of SMOTE to rebalance the dataset. Overall, SMOTE may not have been the best option to balance the data. Another obstacle was cleaning and converting mixed-type columns such as TotalCharges into a usable numerical format without losing data integrity. Lastly, finding the right balance between maximizing model performance and maintaining interpretability was an ongoing consideration throughout model development.

**Future Uses/Additional Applications**

This churn prediction framework can be further developed and applied in real-world business settings. It could be integrated into a company’s operational systems for real-time monitoring and intervention when a customer shows signs of churn risk. Additionally, it could be extended by combining it with customer relationship management (CRM) data to enable personalized retention strategies. The model architecture could also be adapted for deployment across different geographic regions or product lines, assuming similar data is available. However, it is important to note that these models are in preliminary phases and require significant additional testing and adjusting prior to implementation.

**Recommendations**

At this time, I recommend additional models be built and tested to yield more accurate results. I think more data needs to be collected and the class imbalance needs to be addressed in order to implement a model using current real time data. Model builders should segment the data in a way that is accurate for specific regions but captures churn rationale over longer periods of time. With reasoning consistently changing, consistent testing of model accuracy is necessary to deliver the most accurate results to companies on churn predictions.

**Implementation Plan**

Once a more accurate model is developed. It should be implemented using real-time data and tested again to ensure it has consistent accuracy. Each month, the model should be reevaluated to ensure it is producing accurate results In understanding customer churn. As the market shifts, the reasons for churning may shift as well, and it is important to notate these changes.

**Ethical Assessment**

This project does not make use of personally identifiable information, ensuring that customer privacy is preserved throughout the analysis. However, ethical considerations remain critical when developing predictive models. Care must be taken to avoid introducing or reinforcing bias related to demographic attributes such as age, gender, or income. These protected characteristics should be monitored and, where appropriate, excluded from influencing predictions.

**Works Cited**

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and

mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy, 30(10–11), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson Education.

Verbeke, W., Martens, D., & Baesens, B. (2012). Social network analysis for customer churn

prediction. Applied Soft Computing, 14, 431–446. https://doi.org/10.1016/j.asoc.2013.09.017

**Audience Questions**

1. Why did you choose logistic regression as one of the models for a classification problem?
2. Logistic regression is a widely used and interpretable model for binary classification problems like customer churn. It allows us to establish a clear relationship between the probability of churn and various predictor variables while offering insight into feature coefficients.
3. What steps did you take to address the class imbalance in the churn variable?
   1. The dataset showed a significant imbalance between customers who churned and those who did not. To address this, I used SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic examples of the minority class in the training data. This helped balance the class distribution and allowed the models to learn patterns related to churn more effectively.
4. What were the most important features across all 3 models that contributed to results?
   1. Across Logistic Regression, Random Forest, and XGBoost, the most important features influencing churn included contract type, tenure, and monthly charges. Customers on short-term or month-to-month contracts, with lower tenure and higher monthly charges, showed a higher likelihood of churn. Additionally, payment method and paperless billing were also relevant indicators.
5. If you had more time or resources, how would you improve the model’s performance?
   1. With more time, I would explore additional models. Hyperparameter tuning using grid search or Bayesian optimization could also improve performance. Furthermore, adding time-based data spanning over longer periods, customer engagement metrics, or sentiment analysis from support interactions would likely yield deeper insights into churn behavior.
6. How could this model be integrated into a real-time system for churn prediction?
   1. The model can be deployed in a customer database. As new customer activity data flows in, the system can evaluate churn risk in real time and trigger workflows for proactive retention strategies.
7. What are the risks of false positives across each model result and how should the business handle them?
   1. A false positive occurs when the model predicts a customer will churn when they won’t. While not as damaging as a false negative, it may lead to unnecessary retention offers, increasing costs. To mitigate this, the business could prioritize intervention based on confidence scores, applying more aggressive offers only to customers with high predicted risk and high value.
8. What additional features would you look at in addition to the ones within this dataset to help understand customer decisions?
   1. I would look more into features like customer satisfaction and support tickets to see the overall behavioral and sentiment data for the company.
9. How do you ensure that each model remains fair and unbiased?
   1. It is crucial to monitor model performance against demographic segments to avoid bias and keep the data fair.
10. What processing steps did you take to complete your exploratory data analysis and handling inconsistencies within the data?
    1. This was a pretty clean dataset as I found no missing or duplicate values. I removed unnecessary columns and converted categorical data into dummy variables. I used SMOTE to handle class imbalance and built 3 models.
11. How scalable is this solution for a telecommunications provider that has millions of customers and real-time data flows?
    1. This solution is only a small representation of a global company. I would continue to analyze models on larger batches of data over longer periods of time. It may be beneficial to include big data frameworks such as Apache Spark and Hive to build a larger model that can process information at larger and speedier capacities.