Customer Churn

Westyn Hilliard and Matthew Garlock

Customer Churn

Customer churn is a significant challenge for subscription-based businesses, impacting revenue and profitability. High churn rates can lead to substantial financial losses because acquiring new customers is typically more costly than retaining existing ones (McCarthy & Bogdan, 2022). By predicting customer churn, businesses can proactively develop strategies to retain customers before they leave. This problem is crucial for several stakeholders, including customer relationship management (CRM) teams focused on retention, marketing departments dedicated to engagement and loyalty, and business executives interested in maximizing profitability by reducing churn. This project provides a tool for identifying at-risk customers, enabling targeted retention actions that can help reduce churn and improve long-term profitability.

For this analysis, we used the Telco Customer Churn Dataset from Kaggle, which includes essential features such as customer demographics, account details, service usage, and billing information (BlastChar, 2018). These features offer a comprehensive foundation for analyzing and predicting churn behavior. The dataset's variety of features allows us to understand different aspects of customer behavior, making it possible to answer questions about why customers churn and which factors predict churn likelihood most.

Exploring the data provided several insights. First, the churn distribution shows a significant imbalance between churned and non-churned customers, indicating the need for class imbalance handling in the modeling process. Key features such as contract type, internet service, and payment method were also found to influence churn. For instance, customers with month-to-month contracts are likelier to churn, while those on one-year or two-year contracts show lower churn rates. A correlation analysis revealed that features like tenure negatively correlate with churn, suggesting that customers with longer relationships with the company are less likely to leave. We used visualizations to communicate these insights effectively, including churn distribution, feature distributions, and categorical feature analyses. These visualizations provide a straightforward story of how different customer attributes relate to churn, making it easier for stakeholders to understand the data.

To prepare the data, we cleaned it by converting `Total Charges` to a numeric format and removing missing values. We also encoded categorical variables, such as `gender` and `Partner`, to ensure compatibility with modeling techniques. We split the data into training and test sets to validate our models. We used Logistic Regression, Random Forest, and Gradient Boosting Machine (GBM) for modeling. Logistic Regression served as a baseline due to its simplicity and interpretability, while Random Forest and GBM were chosen for their ability to capture non-linear relationships and feature interactions (Chen & Guestrin, 2016). The evaluation metrics included accuracy, precision, recall, F1-score, and AUC-ROC, which are particularly suited to our imbalanced dataset. Metrics like precision and recall helped us understand the model's ability to identify churners accurately, while AUC-ROC provided a holistic view of each model’s discriminatory power. Given the class imbalance, these metrics were more informative than accuracy alone, enabling us to assess the model’s performance more comprehensively.

This project taught us that contract type, payment method, and tenure significantly influence customer churn. Customers on month-to-month contracts and those paying via electronic checks are at higher risk of churning, highlighting areas where the company could focus its retention efforts. Based on these findings, we recommend that the company implement strategies to encourage longer-term contracts and investigate possible dissatisfaction among customers who use electronic checks (Verbeke et al., 2011). These steps could reduce churn rates and improve customer retention.

Our models, notably the Gradient Boosting Machine, can identify churn patterns effectively, though further tuning and testing could enhance their predictive accuracy. While the model shows promise for deployment, refining the features and exploring additional algorithms could improve its performance. Looking forward, feature engineering (such as interaction terms) and more in-depth analysis of high-risk groups could further strengthen our results.

Regarding ethical considerations, it is essential to ensure that our model does not introduce bias, particularly across demographic groups like gender and age. We must also be vigilant about data privacy, ensuring compliance with regulations like GDPR by anonymizing customer data and implementing secure data handling practices. Regular audits of the model for bias and strict adherence to privacy standards are essential to mitigate ethical concerns. If implemented in a production environment, these ethical safeguards would help ensure that the model is fair, transparent, and respectful of customer privacy.

In summary, this analysis demonstrates the potential of using data-driven insights to reduce customer churn, with implications for improved customer retention and profitability. By addressing these recommendations and ethical considerations, the company can enhance its customer experience while operating responsibly and privacy-consciously.

References

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

McCarthy, J., & Bogdan, R. (2022). *Customer Retention Strategies in the Telecom Industry*. Journal of Business Research, 115, 123-134. https://doi.org/10.1016/j.jbusres.2021.08.022

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). *Building comprehensible customer churn prediction models with advanced rule induction techniques*. Expert Systems with Applications, 38(3), 2354-2364. https://doi.org/10.1016/j.eswa.2010.08.023