CUSTOMER CHURN ANALYSIS IN THE TELECOM INDUSTRY

Submitted by

ANUBHUTI ANURAG (23FE10CSE00842)

Under the Guidance of

MR. ASHOK KUMAR SAINI

Assistant Professor, Department of Computer Science and Engineering

In partial fulfillment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
MANIPAL UNIVERSITY JAIPUR
RAJASTHAN 303007

JULY 2025

Introduction

In the rapidly evolving and highly competitive telecom sector, retaining existing customers has emerged as a critical challenge. The cost incurred in acquiring new customers typically outweighs the expenses involved in nurturing and retaining the current customer base. Customer churn, which denotes the phenomenon of customers discontinuing their services, poses a direct threat to the profitability and long-term sustainability of telecom enterprises. This project is centered around addressing this pressing issue by employing a data-driven framework. Through the use of advanced analytics and machine learning models, it seeks to predict churn with substantial accuracy and offer actionable strategies aimed at improving customer retention. By systematically analyzing diverse customer data, the project identifies behavioral trends and risk factors associated with churn, ultimately empowering telecom providers to proactively engage with customers who are most likely to leave. This contributes significantly to enhancing customer lifetime value and securing a competitive edge in the marketplace.

Objective

The principal objective of this project is to construct an effective churn prediction system capable of accurately flagging customers who are on the verge of terminating their services. This has been achieved by leveraging a binary classification machine learning approach that utilizes the IBM Telco Customer Churn dataset. Alongside predicting churn, the project aims to unravel the pivotal indicators that drive customer attrition and translate these findings into insightful strategies for the business teams. It is designed not merely to forecast churn but also to facilitate intelligent segmentation of the customer base, thereby recommending tailored retention initiatives for each segment. Furthermore, the project aspires to streamline organizational decision-making by integrating automated churn scoring mechanisms within Customer Relationship Management (CRM) systems, thereby bolstering customer loyalty through targeted and reward-based engagement campaigns.

To ensure high model performance, extensive data preprocessing techniques such as feature encoding, handling class imbalance, and outlier treatment were implemented. Various classification algorithms, including Logistic Regression, Random Forest, and XGBoost, were evaluated to identify the most robust and interpretable model. The selected model was further fine-tuned using hyperparameter optimization techniques like Grid Search and Cross-Validation. Model explainability was enhanced using SHAP (SHapley Additive exPlanations) values to provide transparency into feature contributions. A dynamic dashboard was also developed to visualize key churn metrics,

customer segments, and actionable insights for non-technical stakeholders. The integration pipeline was designed with scalability in mind, ensuring seamless deployment across enterprise-level infrastructures. Ultimately, this project represents a strategic blend of data science and business intelligence, offering a comprehensive solution to proactively manage and reduce customer churn.

Motivation

The driving force behind this project stems from the urgent need within the telecom industry to manage customer retention more intelligently and efficiently. Telecom companies frequently face substantial revenue setbacks due to unforeseen customer attrition, which is often exacerbated by an inadequate grasp of customer preferences and insufficient exploitation of the wealth of available data. The advent of sophisticated machine learning techniques and expansive data analytics capabilities presents a promising avenue to address this issue. It enables organizations to anticipate churn well in advance and implement strategic measures to mitigate it. This project endeavors to harness these technological advancements by designing a practical and robust churn prediction framework that equips the business to make proactive and well-informed decisions. Ultimately, it seeks to foster greater customer satisfaction, minimize churn rates, and elevate the overall business performance through judicious data utilization.

To achieve this, the project emphasizes a data-driven culture where customer behavior patterns are continuously monitored and interpreted. It also aims to bridge the gap between technical models and business applicability through interpretable results. Furthermore, the framework is structured to evolve with changing market dynamics, ensuring its long-term relevance. By aligning technical innovation with customer-centric strategies, this initiative aspires to become a cornerstone of sustainable competitive advantage in the telecom sector.

Methodology

The methodology adopted for this churn analysis project commenced with comprehensive data acquisition, relying on the IBM Telco Customer Churn dataset, which encompasses over 7,000 customer records enriched with demographic profiles, account specifics, service usage details, and churn status labels. A meticulous data cleaning phase followed, addressing missing values notably present in the TotalCharges attribute by converting blanks into valid numerical entries, while extraneous features like customerID were eliminated to enhance the dataset’s predictive quality.

Categorical variables were systematically transformed through label encoding to prepare them for machine learning algorithms. Given the inherent class imbalance, where churners constituted merely 26.5% of the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed to create a balanced class distribution. Exploratory Data Analysis (EDA) played a pivotal role in this phase, uncovering significant correlations between churn tendencies and variables such as tenure duration, contract type, payment method, and technical support interactions. Feature relationships were meticulously visualized using histograms, kernel density plots, and correlation matrices. The modeling stage incorporated Decision Tree, Random Forest, and XGBoost classifiers, with Random Forest emerging as the most balanced model, attaining an accuracy close to 84% coupled with a commendable trade-off between precision and recall. The interpretability of the model was further enhanced using SHAP analysis to elucidate the contribution of each feature toward the prediction. In parallel, SQL queries were crafted to execute aggregate analysis, revealing churn patterns across various demographic and service usage dimensions, thereby reinforcing the analytical foundation of the solution.

ER Diagram

The Entity-Relationship diagram crafted for this customer churn analysis project has been meticulously tailored to reflect the multifaceted dynamics of customer interactions that contribute to churn in the telecom context. At the core lies the Customer entity, which is uniquely identified by customer attributes such as demographic details and account identifiers. This entity is directly linked to the Subscription entity, which encapsulates details like contract type, tenure period, and subscribed services that play a vital role in influencing retention. The Payment entity maintains records of billing preferences, monthly charges, total accumulated charges, and payment methods, offering insights into customers' financial engagement and potential billing-related triggers for churn. Additionally, the Service Interaction entity captures logs of technical support calls, issue resolutions, and requests for upgrades or downgrades, all of which serve as indicators of customer satisfaction or frustration. These entities are interconnected through well-defined relationships that establish a comprehensive view of how various factors, ranging from service quality to billing experience, converge to affect the likelihood of churn. This diagram serves as a conceptual backbone for the database design and is instrumental in driving SQL-based analytical queries as well as structuring features for machine learning models.

Constraints

Despite delivering robust analytical and predictive capabilities, the project operates under several practical and methodological constraints. A significant limitation stems from the dependence on historical data; the predictive accuracy is inherently tied to the quality, granularity, and completeness of past customer records. Any inconsistencies, outdated fields, or missing values can diminish the model's reliability and the relevance of insights drawn. The dataset utilized, derived from the IBM Telco Customer Churn records, represents specific market behaviors and customer profiles that may not universally mirror patterns across all telecom providers, thereby restricting the generalizability of the developed model. Additionally, customer preferences and competitive dynamics in the telecom industry evolve rapidly; hence, static models might lose efficacy over time without periodic retraining or updates to incorporate fresh data. Another notable constraint lies in the interpretability of complex ensemble techniques such as Random Forests and XGBoost. Although tools like SHAP have been employed to decode feature contributions, business stakeholders in regulatory or compliance-heavy environments might still demand simpler, transparent decision rules that are harder to extract from these models. Furthermore, practical deployment in production environments could introduce constraints tied to real-time data integration, latency, and system scalability, which would necessitate additional engineering considerations beyond the current analytical scope.

SQL Queries

To deepen the understanding of churn patterns beyond predictive modeling, a suite of SQL queries was designed to distill actionable aggregates and segment-wise insights from the underlying relational database. These queries encompass calculations of churn rates segmented by contract types, which reveal that customers on flexible month-to-month plans exhibit substantially higher churn probabilities compared to those committed to longer two-year contracts, likely due to lower switching costs. Another set of queries explores the influence of payment methods, highlighting, for instance, how customers who opt for electronic check payments show a greater tendency to churn, possibly due to perceived complexity or dissatisfaction with automated billing processes. Additional analytical queries delve into the tenure brackets, quantifying how short-tenure customers (new sign-ups) are considerably more prone to churn, emphasizing the critical window for engagement in the early customer lifecycle. Queries have also been executed to correlate churn with paperless billing adoption and the frequency of technical support interactions, shedding light on the paradox where higher support calls sometimes indicate underlying service dissatisfaction leading to churn.

Collectively, these SQL-driven insights complement the machine learning outputs, furnishing the business with a holistic perspective that blends statistical rigor with direct operational relevance.

Outputs

The culmination of this project is encapsulated in a comprehensive suite of outputs that together form an integrated churn mitigation toolkit. Central to this is a predictive classification engine, which processes customer profiles and assigns churn probabilities with a significant degree of confidence. This engine is paired with interactive dashboards that visualize the distribution of churn risk across the customer base, highlight the most influential features contributing to churn, and segment customers into actionable risk tiers such as low, medium, and high churn propensity groups. These dashboards enable marketing and retention teams to swiftly identify vulnerable segments and tailor interventions. The project further produces SQL-based summary tables that quantify churn rates across demographic slices and service usage patterns, which are instrumental for strategic planning and resource allocation. SHAP value plots accompany the outputs to provide transparent explanations of individual predictions, fostering trust in automated decisions. Taken together, these outputs transform complex data and sophisticated modeling results into accessible, business-friendly intelligence, empowering decision-makers to launch proactive campaigns—whether through loyalty rewards, personalized offers, or direct engagement—to substantially curb churn.

Conclusion

This project stands as a testament to the power of data-driven strategies in addressing one of the telecom industry's most challenging hurdles: customer churn. By seamlessly integrating rigorous data cleaning, exploratory data analysis, and advanced machine learning algorithms, the initiative has succeeded in constructing a predictive system that not only forecasts churn with appreciable precision but also unpacks the multifactorial reasons behind customer attrition. The adoption of interpretability frameworks such as SHAP has ensured that model decisions are not black boxes but are instead accompanied by clear, feature-level justifications, thus bridging the gap between technical outputs and strategic business actions. The insights drawn empower telecom providers to transition from reactive churn management to a proactive stance, where at-risk customers are identified and engaged before they exit. Looking ahead, the foundation laid by this project could be expanded to incorporate real-time data pipelines, enabling continuous monitoring and instant intervention capabilities, thereby evolving into a dynamic and responsive churn prevention ecosystem that adapts fluidly to market and customer changes.

Facilities Used

The successful realization of this project was made possible through the deployment of a diverse set of computational tools and platforms. The core analytical and modeling tasks were executed in Python, leveraging libraries such as Pandas for data manipulation, NumPy for numerical computations, Matplotlib and Seaborn for intricate visualizations, Scikit-learn for preprocessing and machine learning workflows, XGBoost for powerful gradient boosting models, and SHAP for detailed interpretability analysis. SQL Server played a crucial role in handling structured data operations and executing complex aggregation queries that provided pivotal exploratory insights. The project also benefited from the use of Jupyter Notebooks, which facilitated an interactive coding environment ideal for iterative development and visualization. Hardware resources comprised a modern high-performance workstation equipped with multi-core processors and ample RAM, ensuring that even computationally intensive tasks such as SMOTE balancing and cross-validation ran efficiently. These facilities collectively provided a robust environment for orchestrating the end-to-end churn analysis pipeline, from initial data ingestion to the delivery of final business-ready insights.

References

1. IBM Telco Customer Churn Dataset – Available on Kaggle and originally from IBM Watson Analytics.
2. Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. Book link
3. Molnar, Christoph. *Interpretable Machine Learning*. Online book
4. Pandas library documentation – https://pandas.pydata.org/docs/
5. Scikit-learn library documentation – https://scikit-learn.org/stable/user_guide.html
6. XGBoost library documentation – https://xgboost.readthedocs.io/en/stable/
7. McKinsey & Company. *How telecom companies can win in the customer retention game*. Article link