Customer Churn Prediction Project: Driving Retention with Qlik Analytics

Project Goal: To identify customers at high risk of churning and provide actionable insights to a telecommunications company, enabling proactive retention strategies and minimizing revenue loss.

Customer churn is a critical challenge for subscription-based businesses, directly impacting revenue and growth. Understanding why customers leave and who is likely to leave allows companies to intervene effectively. This project demonstrates an end-to-end analytical approach to tackle this business problem.

Data Source & Preparation

The project utilized a comprehensive Telecom Customer Churn Dataset obtained from Kaggle (Attached below), containing various customer attributes, service usage patterns, and billing information. After downloading from Kaggle I looked over the dataset realizing that the data quality all in all was pretty good. However, I saw some areas that I felt needed some cleaning. I also had a few ideas about new columns I could add to the dataset. I fired up Gemini and requested the Gen AI to convert the csv file I had downloaded into a SQL script that I could then use in MySQL Workbench to clean up the file. I also used Gemini and its fellow LLM Chat GPT to generate the apply datasets which were used for predictions after the Qlik ML model had been trained on the original telecom churn dataset. These apply datasets consisted of all the same columns as the original that was used to train the model, however this time the data in all the columns but the churn column (intentionally left blank for the model to predict) was randomized data that fit the structure of the original database. The first apply dataset was used as a trial run with only 500 rows of data before moving onto a larger dataset where this time I asked Gemini to generate an apply data set with between 5,000 to 7,000 rows of new data. Gemini kindly obliged and delivered me a brand new apply dataset with 6,858 rows which was then fed into the prediction model to provide the new churn prediction data used to develop the visualizations in the deliverables section. The final area in which I utilized Gen AI was to help with the initial SQL queries exclusively the DECIMAL, REPLACE, and REGEXP queries

virtually all of which were within just the first 3-4 set up queries, the remainder of the SQL queries were written without the assistance of LLM's.

Data Cleaning & Feature Engineering in MySQL: Once I had my SQL script loaded into MySQL I then began querying, cleaning, and the creation of new, insightful features directly within a MySQL database environment. This ensured the dataset was optimized for predictive modeling and provided richer analytical dimensions.

Initial Cleaning and Type Conversion of Total Charges: The Total Charges column initially contained non-numeric values (empty strings), preventing direct numerical analysis. To address this, all empty string values (and strings with only spaces, and NULLs) in Total Charges were updated to '0'. Any remaining non-numeric characters (like '$', ',') were removed. Finally, the Total Charges column was successfully converted to a DECIMAL (10, 2) data type, allowing for accurate mathematical operations.

Engineering New Features: I added four new features to the telco_customer_churn table to enhance the predictive power of the model:

Tenure Months Binned: The tenure column (in months) was transformed into categorical bins: '0-12 Months', '13-24 Months', '25-48 Months', '49-72 Months', and '72+ Months'. This allowed for analyzing churn patterns across distinct customer lifecycle stages.

Total Service Addons: An integer column quantifying the number of additional services (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies) a customer subscribed to. This feature helps assess customer engagement.

Has Paperless Electronic Payment: A binary indicator (0 or 1, later converted to 'No'/'Yes') flagging customers who use both 'PaperlessBilling' and 'Electronic check' as their PaymentMethod. This specific combination was hypothesized to be linked to higher churn rates.

Average Monthly Charge Per Tenure: A continuous numerical feature calculating the customer's average monthly spending over their entire tenure (TotalCharges / tenure). For new customers with tenure = 0, their MonthlyCharges value was used to avoid division by zero. This metric provides insight into consistent financial engagement. The column was set to DECIMAL (15, 6) for precision.

Standardizing Categorical Values: For consistency and readability, the Senior Citizen and Has Paperless Electronic Payment columns, which contained '0' and '1' values, were converted to 'No' and 'Yes' respectively, and their data types were updated to VARCHAR (3).

Column Renaming: Finally, the newly engineered columns were renamed to remove underscores, aligning with the naming of the columns in the rest of the table for easier use in analytics platforms (e.g., Tenure_Months_Binned became TenureMonthsBinned).

Final SQL Queries Used: The complete and final SQL statements used for these transformations are provided in the project's GitHub repository.

Predictive Modeling with Qlik ML Prediction

Leveraging Qlik Cloud Analytics' ML experiment and prediction capabilities, I developed a binary classification model to predict customer churn. Qlik ML streamlines the machine learning process, allowing for rapid model development and deployment without extensive coding (or really any coding for that matter), which is ideal for an analyst focused on business outcomes.

The platform automatically explored and trained multiple algorithms, including CatBoost Classification, LightGBM Classification, XGBoost Classification, Random Forest Classification, and Logistic Regression, to identify the best-performing model. The dataset was split into training and a stratified holdout set (20%) to ensure unbiased evaluation of the model's performance on unseen data.

Key Findings & Model Performance

The CatBoost Classification model (v01_CATBC_01_02) emerged as the top performer on the holdout dataset, demonstrating strong predictive capabilities.

Key Performance Metrics (on Holdout Data):

F1-Score: 0.613 (A balanced measure of precision and recall, indicating a good trade-off between identifying actual churners and minimizing false positives.)

AUC (Area Under the Curve): 0.827 (A strong indicator of the model's ability to distinguish between churning and non-churning customers across various thresholds.)

Accuracy: 0.749

Recall: 0.749 (The model successfully identified approximately 75% of actual churners.)

Precision: 0.519 (Approximately 52% of customers predicted to churn did.)

Top Predictive Features: Qlik ML provides critical insights into the features most influential in predicting churn. The most impactful factors were:

Contract Type: This was the most significant predictor, highlighting that customers on month-to-month contracts are significantly more prone to churn compared to those on longer-term agreements.

Tenure: The length of time a customer has been with the service plays a substantial role, often indicating loyalty or early-stage risk.

Internet Service Type: The specific internet service package a customer subscribes to also strongly influences their likelihood of churning.

Total Charges: The cumulative amount billed to a customer over their tenure.

Actionable Insights & Recommendations

Based on the model's insights, the following actionable recommendations could be made to mitigate churn:

Targeted Retention Programs for Month-to-month Customers: Develop specific incentives, loyalty/reward programs, focused marketing and promotional efforts. (e.g., discounts for signing a 1-year contract, exclusive service upgrades when committing to a longer-term deal, etc.) to encourage month-to-month customers to commit to longer-term plans.

Early Intervention for New Customers: Monitor new customers closely, especially those within their first few months of service, and proactively address any potential issues or offer onboarding support to improve early retention. A strong balance must be found between being over attentive to new customers on short term deals as to not lose out on those who have remained loyal customers for a long period of time, therefore shifting the risk of churn to those who have been subscribed for longer.

Service Quality Review by Internet Type: Increase resources to investigate potential service quality issues or dissatisfaction among customers with specific internet service types and implement improvements or offer alternative solutions.

Value-Based Offers for High-Spending Customers: For customers with high total charges, ensure they perceive adequate value for their spending. Consider personalized

offers or loyalty benefits to reinforce their commitment and reward their loyalty to the company as tenure appeared to be the biggest influence on churn.

Dashboards & Visualizations

To make these insights accessible and actionable, I developed interactive dashboards in Qlik Sense. These dashboards allow users to explore predicted churn patterns, identify at-risk customer segments, and understand the drivers of predicted churn visually.

Qlik Dashboard

The dashboard features several key visualizations:

Predicted Cancellations & Churn Rate (KPIs): These Key Performance Indicators provide an immediate executive summary. They show the total number of customers predicted to churn out of the total 6k+ customers and the overall predicted churn rate, offering a quick snapshot of the scale of the retention challenge.

Impact of Contract Type on Predicted Churn (Stacked Bar Chart): This chart vividly illustrates how different contract types correlate with predicted churn. By showing the proportion of predicted churners within each contract term (Month-to-month, one year, two year), it highlights that customers on month-to-month contracts represent the largest segment of predicted churn, emphasizing where retention efforts should be prioritized. When selecting only the month-to-month bar in the interactive dashboard you will realize the churn rate jumps tremendously to just above **32%**!

Tenure's Influence on Churn Risk (SHAP Analysis) (Scatter Plot with SHAP Values): This visualization delves into the individual impact of customer tenure on churn predictions. Each point represents a customer, with their tenure on the X-axis and the tenure's SHAP value (impact on churn prediction) on the Y-axis. It clearly shows that lower tenure (new customers) positively contributes to churn likelihood, while higher tenure generally reduces it, guiding lifecycle-based retention strategies. The text box conveniently located underneath provides a clear explanation of SHAP values for easy interpretation.

Distribution of Predicted Churn Across Total Charges (Area Chart): This chart displays the density of predicted churners and non-churners across various ranges of Total Charges. It helps identify specific cumulative spending tiers where churn is more prevalent, such as the early stages of a customer's journey characterized by low total

charges. This insight can inform strategies related to initial customer value perception and engagement.

Future Predictions & Operationalization

The trained and deployed Qlik ML model can be used to generate real-time or batch predictions on new customer data, allowing a company to identify at-risk customers before they churn. This proactive approach enables timely interventions and resource allocation.

This model could be used in ongoing forecasting or predictions, such as with enough training becoming integrated into daily operations to flag customers whose data may suggest that they are at risk of churning and an ideal course of action may be to include them in a customer retention focused program/marketing ploy by the company. Another use could be general weekly/monthly reports such as reports that list the top 100 customers at the highest risk of churn.

Conclusion

This Customer Churn Prediction project demonstrates my ability to leverage database querying tools and languages such as SQL and advanced analytics tools like Qlik ML to solve real-world business problems. By transforming raw data into predictive insights and actionable recommendations.