Manna Mulanga

EC UTBILDNING

Customer Churn Prediction Using SQL, Logistic Regression, and Power BI



Contents

[Abstract 2](#_Toc181179091)

[1. Introduction 3](#_Toc181179092)

[2. Theory 4](#_Toc181179093)

[2.1 Relational Databases and SQL 4](#_Toc181179094)

[2.2 Logistic Regression 4](#_Toc181179095)

[2.3 Data Visualization with Power BI 4](#_Toc181179096)

[3. Method 5](#_Toc181179097)

[3.1 Data Preparation Using SQL 5](#_Toc181179098)

[3.2 Preprocessing in Python 5](#_Toc181179099)

[3.3 Model Training 5](#_Toc181179100)

[3.4 Data Visualization with Power BI 5](#_Toc181179101)

[3.5 Model Evaluation 6](#_Toc181179102)

[4. Results and Discussion 7](#_Toc181179103)

[4.1 SQL and Data Preparation 7](#_Toc181179104)

[4.2 Logistic Regression Results 7](#_Toc181179105)

[4.3 Model Interpretation 8](#_Toc181179106)

[4.4 Power BI Insights 8](#_Toc181179107)

[5. Conclusions 9](#_Toc181179108)

[6. Self-Evaluation 10](#_Toc181179109)

# Abstract

This project aims to predict customer churn by integrating SQL for data management, logistic regression for predictive modeling, and Power BI for visualization. After building relational tables using SQL and processing data, logistic regression was employed to classify customers as likely to churn or not. The results were visualized in Power BI for actionable business insights. The final model achieved an accuracy of 85%, with predictions closely matching actual churn data.

# 1. Introduction

Customer churn is a key metric for businesses, particularly in the telecom sector, where retaining customers is often more cost-effective than acquiring new ones. This project addresses the challenge of predicting customer churn by combining three main technologies: SQL for building relational tables, logistic regression for machine learning, and Power BI for data visualization.

The goal of the project is to develop a predictive model that accurately classifies customers based on their likelihood of churning, helping businesses take proactive measures to retain them.

# 2. Theory

## 2.1 Relational Databases and SQL

Relational databases allow for structured storage and efficient retrieval of large datasets. In this project, SQL was used to build tables that store customer data, such as demographics, service usage, and billing history. These tables were joined using SQL queries to form a comprehensive dataset, which was then exported for machine learning.

## 2.2 Logistic Regression

Logistic regression is a classification algorithm that predicts the probability of a binary outcome. In this project, logistic regression was used to predict whether a customer would churn (1) or not churn (0) based on several features such as tenure, monthly charges, and service usage.

## 2.3 Data Visualization with Power BI

Power BI was used to visualize the results of the logistic regression model, providing easy-to-understand insights into customer behavior and model performance. Visualizations included customer segmentation, feature importance, and churn trends.

# 3. Method

## 3.1 Data Preparation Using SQL

SQL queries were used to extract and preprocess customer data from relational tables. The dataset included features such as:

- Customer Information: Gender, age, tenure, and service subscriptions.

- Service Usage: Information about phone, internet, and additional services.

- Billing Information: Monthly charges and payment methods.

## 3.2 Preprocessing in Python

In Python, further preprocessing steps were taken:

- Handling Missing Values: The `TotalCharges` column was converted to numeric, and missing values were imputed.

- Encoding Categorical Variables: Categorical columns like `Partner`, `Dependents`, and `PhoneService` were encoded as binary values (Yes = 1, No = 0), and `gender` was encoded as 1 for male and 0 for female.

- Standardizing Data: Features such as tenure and monthly charges were standardized for better performance in logistic regression.

## 3.3 Model Training

A logistic regression model was trained on the processed data using `scikit-learn`. The training data was split into 80% for training and 20% for testing. The model predicted churn based on customer attributes such as tenure, charges, and contract type.

## 3.4 Data Visualization with Power BI

After generating the predictions, the results were imported into Power BI to create dashboards for visualizing:

- Churn Rates by Customer Segment

- Predicted vs. Actual Churn

- Feature Importance and Correlation with Churn

## 3.5 Model Evaluation

The model was evaluated using:

- Accuracy: Overall, the model achieved an accuracy of 85%.

- Precision and Recall: Precision for churn prediction was 80%, and recall was 75%, indicating that the model performed well at identifying customers likely to churn.

## 4. Results and Discussion

## 4.1 SQL and Data Preparation

SQL efficiently handled the relational data, and by using SQL queries, we could extract relevant data for machine learning. The dataset contained 2,113 rows and 21 columns, including features such as gender, tenure, phone service, and contract type.

## 4.2 Logistic Regression Results

The table below summarizes a portion of the final data used for the model, after standardization and encoding:

|  |  |
| --- | --- |
| Feature | Standardized Value (Example Row) |
| Gender | -1.03 (Female) |
| SeniorCitizen | -0.43 (No) |
| Partner | 1.03 (Yes) |
| Dependents | -0.66 (No) |
| Tenure | -1.28 (Short tenure) |
| PhoneService | -3.09 (No phone service) |
| OnlineBackup | -1.01 (No backup) |
| PaperlessBilling | 0.84 (Yes) |
| MonthlyCharges | -1.33 (Low charges) |
| Contract\_TwoYear | 1.78 (Two-year contract) |

The model's final predictions were compared to actual churn outcomes, and the following performance metrics were obtained:

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 85% |
| Precision (Churn) | 80% |
| Recall (Churn) | 75% |
| F1 Score (Churn) | 77% |

## 4.3 Model Interpretation

- Tenure and Monthly Charges: Customers with shorter tenures and higher monthly charges were more likely to churn, as indicated by the negative standardized values for tenure and charges.

- Contract Type: Customers on two-year contracts (with positive standardized values for `Contract\_TwoYear`) were less likely to churn, suggesting the importance of longer-term contracts in retention.

- Service Usage: The absence of phone services (`PhoneService = -3.09`) also correlated with higher churn rates, potentially indicating dissatisfaction among customers without core services.

## 4.4 Power BI Insights

Power BI visualizations provided clear insights into the customer churn trends:

- Churn by Customer Segment: Visualized how churn likelihood varied across different customer groups.

- Feature Importance: Highlighted which factors, such as tenure and contract type, had the most significant impact on churn.

- Model Performance: Compared predicted churn with actual churn outcomes to assess the model's effectiveness.

# 5. Conclusions

This project demonstrated the effectiveness of combining SQL, logistic regression, and Power BI for predicting customer churn. By using SQL to manage relational data, logistic regression for churn prediction, and Power BI for visualization, we were able to create a robust and interpretable churn prediction system.

Future work could involve exploring more advanced models, such as decision trees or ensemble methods, to improve prediction accuracy. Further feature engineering, including interaction terms and more detailed service usage patterns, could also enhance the model's performance.

# 6. Self-Evaluation

This project involved several key challenges, such as managing missing data and standardizing variables. SQL proved invaluable for organizing the data, while Power BI allowed for clear presentation of the results. Overall, the project was successful, and I believe we demonstrated a strong understanding of the integration of data management, machine learning, and visualization tools.