B2B SaaS Churn Prediction Project

Overview

This project focuses on predicting customer churn for a B2B SaaS company using machine learning techniques and visualizing actionable insights in Power BI. The goal is to identify at-risk customers, understand the drivers of churn, and provide recommendations to reduce churn and increase revenue.

The primary objective of this project is to:

Predict customer churn using machine learning models.

Identify key factors contributing to churn.

Provide actionable insights to reduce churn and improve customer retention.

Visualize churn-related metrics and trends for business stakeholders.

Dataset and Data Sources

**The dataset was generated and processed to simulate real-world B2B SaaS customer behavior. It includes the following key columns:**

Customer Details

CUSTOMER\_ID: Unique identifier for each customer.

SEGMENT: Customer segment (e.g., SMB, Mid-Market, Enterprise).

INDUSTRY: Industry of the customer.

COUNTRY: Customer's country.

Subscription Details

PLAN: Subscription plan (e.g., Basic, Premium, Enterprise).

SUBSCRIPTION\_DURATION: Duration of the subscription in months.

MRR: Monthly Recurring Revenue.

Usage Metrics

AVG\_MONTHLY\_LOGINS: Average monthly logins by the customer.

AVG\_TICKETS\_RAISED: Average number of support tickets raised.

AVG\_FEATURE\_USAGE: Average feature usage score.

Churn Prediction

ACTUAL\_CHURN: Whether the customer churned (1) or not (0).

PREDICTED\_CHURN: Model's prediction of churn.

CHURN\_PROBABILITY: Probability of churn predicted by the model.

RISK\_LEVEL: Categorized churn risk (Low, Medium, High).

REVENUE\_AT\_RISK: Revenue at risk due to potential churn.

Project Workflow

The project was executed in the following steps:

1. Data Generation

Synthetic data was generated using Python libraries (pandas, numpy, random) to simulate customer behavior.

The data was saved into CSV files:

customer\_accounts.csv

subscriptions.csv

usage\_metrics.csv

2. Data Upload to Snowflake

The generated CSV files were uploaded to Snowflake using:

Snowflake Web Interface: For staging and loading data.

SQL Queries: To create tables and load data into Snowflake.

3. Feature Engineering

SQL queries were used to engineer features such as:

Subscription duration.

Revenue at risk.

Aggregated usage metrics.

Null values were handled, and active subscriptions were calculated using the current date.

4. Modeling

The feature-engineered dataset was exported from Snowflake and loaded into a Jupyter Notebook.

Steps:

Data Preprocessing:

Handled missing values.

Encoded categorical variables.

Scaled numerical features.

Class Imbalance Handling:

Used SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset.

Model Training:

Trained multiple models, including Logistic Regression, Random Forest, and XGBoost.

Evaluation:

Evaluated models using metrics like accuracy, precision, recall, and F1-score.

5. Visualization

The final dataset, including predictions, was uploaded to Power BI for visualization.

Key metrics and trends were visualized to provide actionable insights.

Tools and Libraries Used

1. Python Libraries

pandas, numpy: Data manipulation and analysis.

scikit-learn: Machine learning and evaluation.

imbalanced-learn: SMOTE for handling class imbalance.

matplotlib, seaborn: Data visualization.

2. Snowflake

SQL for querying, feature engineering, and data export.

3. Power BI

For creating interactive dashboards and visualizations.

Modeling Insights

Best Model: **Random Forest achieved the highest performance with an accuracy of 95%.**

Key Features Driving Churn:

Low feature usage.

Certain geographies.

Low monthly logins.

Class Imbalance: Addressed using SMOTE, which improved model performance by ensuring balanced training data.

Revenue at Risk: Identified high-risk customers contributing to significant revenue loss.

Visualizations and Insights

1. Churn Distribution

Analysis:

The donut chart reveals that 3.28% of customers are predicted to churn, while 96.72% are likely to stay. Although the churn percentage appears small, it represents a significant revenue risk, especially for high-value customers.

This highlights the importance of focusing on the small subset of customers at risk to prevent revenue loss.

Recommendations:

(i) Proactive Retention Campaigns: Launch targeted campaigns for the 3.28% of customers predicted to churn.

(ii) Use personalized offers, discounts, or enhanced support to retain them.

(iii) Customer Segmentation: Further segment the churned customers by industry, geography, and plan to identify patterns and tailor interventions.

(iv) Monitor Trends: Regularly track churn rates to identify any upward trends and act swiftly.

2. Churn by Plan

Analysis:

The bar chart shows that the Growth plan has the highest number of churned customers, followed by the Pro and Enterprise plans.

This suggests that customers on the Growth plan may not be finding sufficient value or may be outgrowing the plan.

Recommendations:

(i) Plan Optimization: Evaluate the Growth plan's features and pricing. Consider adding value through additional features or better support.

(ii) Upsell Opportunities: Identify customers on the Growth plan who are at risk of churn and offer them tailored upsell opportunities to the Pro or Enterprise plans.

(iii) Customer Feedback: Conduct surveys or interviews with Growth plan customers to understand their pain points and address them.

3. Churn by Segment

Analysis:

The SMB (Small and Medium Businesses) segment has the highest churn rate, followed by Mid-Market and Enterprise.

SMBs are often more price-sensitive and may churn due to cost or lack of perceived value.

Recommendations:

(i) Cost-Effective Solutions: Offer SMBs cost-effective solutions or flexible payment plans to reduce churn.

(ii) Value Demonstration: Provide SMBs with case studies, ROI calculators, or success stories to demonstrate the value of the product.

(iii) Dedicated Support: Assign account managers or provide enhanced support to SMBs to build stronger relationships.

4. MRR at Risk by Industry

Analysis:

The Retail and Professional Services industries have the highest MRR at risk, followed by Logistics and Real Estate.

This indicates that these industries may face unique challenges or may not be fully utilizing the product.

Recommendations:

(i) Industry-Specific Solutions: Develop tailored solutions or features for high-risk industries like Retail and Professional Services.

(ii) Customer Success Programs: Assign industry-specific customer success managers to help these customers maximize the value of the product.

(iii) Partnerships: Collaborate with industry associations or thought leaders to build trust and credibility.

5. Total MRR at Risk by Geographies

Analysis:

The Treemap highlights that the USA has the highest MRR at risk, followed by the UK, Brazil, and India.

This suggests that these geographies should be prioritized for churn prevention efforts.

Recommendations:

(i) Localized Strategies: Develop region-specific retention strategies, considering cultural and economic factors.

(ii) Regional Teams: Strengthen regional customer success teams to provide localized support and build stronger relationships.

(iii) Market Research: Conduct market research in high-risk geographies to understand customer needs and address them effectively.

6. MRR at Risk Globally

Analysis:

The global map visualization shows that North America and Europe have the highest concentration of MRR at risk.

This aligns with the Treemap insights and emphasizes the need for a global yet localized approach.

Recommendations:

(i) Global Retention Task Force: Establish a task force to focus on high-risk regions and coordinate efforts across teams.

(ii) Data-Driven Insights: Use predictive analytics to identify early warning signs of churn in these regions and act proactively.

(iii) Customer Advocacy Programs: Build customer advocacy programs in these regions to strengthen loyalty and reduce churn.

Final Dataset: data/final\_churn\_prediction.csv