**1 Identify & Gather Data**

**Customer Demographics Table:**

The Customer Demographics table provides background information on customers, including attributes such as age, gender, marital status, and income level. These demographic factors often correlate with customer retention or churn behaviour. By analysing this data, we can identify segments that are more likely to churn, enabling the bank to design targeted retention strategies and personalized services.

**Transaction History Table:**

This table provides a record of transactions made by each customer. Transaction data serves as an important indicator of customer engagement and financial activity. Patterns such as declining transaction frequency or reduced spending could signal dissatisfaction or disengagement, both of which are closely linked to churn. Aggregating this data allows us to derive key features like total spend, average transaction value, and changes in behaviour over time.

**Customer Service Table:**

Data from this table captures interactions between customers and the bank’s support team. This data is critical, as frequent or unresolved service issues often correlate with a higher likelihood of churn. By analysing interaction frequency, resolution status, and the types of interactions (e.g., complaints vs. inquiries), we can quantify customer satisfaction levels and identify at-risk individuals.

**Online Activity Table:**

This table reflects customers’ engagement with digital services, including metrics like last login date, login frequency, and overall service usage. A decline in digital activity is a common behavioural signal of disengagement. These features serve as strong predictors in churn modelling, helping to detect when a customer is becoming inactive or losing interest.

**Churn Status Table:**

The Churn Status table indicates whether a customer has churned (1) or remained active (0). This is the target variable for our supervised learning model. It allows us to train, validate, and evaluate predictive models that aim to forecast future churn based on the behavioural and demographic characteristics derived from the other tables.

**2 Exploratory Data Analysis**

Exploratory data analysis began by merging all the tables into a single dataset using ‘CustomerID’ as the join key. This allowed for easier comparison of the features and identification of trends.

A screenshot of a computer screen

AI-generated content may be incorrect.**Handling Missing Values:**

After the merge, missing values were assessed. The results initially indicated many missing values in columns from the customer service table. However, upon closer inspection, it was evident that these values are not truly missing. Instead, they reflect customers who have had no recorded interactions with customer service. Because these customers do not appear in the customer service table, the merge operation filled in these fields as missing (NaN). This is an acceptable outcome of the outer join used during merging.

Since there are no actual missing values in the original customer service data, we can proceed with the analysis without additional data cleaning for this issue.

**Numerical Outliers & Distribution:**

A graph of a graph of a graph

AI-generated content may be incorrect.Next, to gain an initial understanding of the data distribution, histograms and boxplots were created for the numerical columns. Histograms provide a clear view of the spread and shape of the data, helping to identify any skewness or irregular patterns. Boxplots are useful for detecting outliers and understanding the overall variation within each feature.

A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a diagram

AI-generated content may be incorrect.The resulting histograms indicate a relatively even distribution of data points across the dataset. While there are some noticeable peaks in variables such as **Age** and **Login Frequency**, these appear to be natural variations rather than anomalies. The boxplots support these observations, showing no significant outliers and indicating that most values are clustered around the centre of the distribution. Overall, there are no immediate concerns with the distribution or spread of the numerical data.

A screenshot of a graph

AI-generated content may be incorrect.

The heatmap above shows how closely related the different features in the dataset are to each other. Based on the results, there don’t appear to be any strong connections between the features. However, this doesn’t mean there’s no relationship at all—it’s possible that some features are connected in ways that aren’t shown by this kind of analysis, such as through non-linear patterns or through categories rather than numbers.

**Categorical Outliers & Distribution:**

A graph of a customer demographics distribution of gender

AI-generated content may be incorrect.Now that the numerical data has been explored, the next step is to look at the categorical variables. These are different because they represent groups or labels rather than numbers. Instead of using box plots, which are suited for numerical data, bar charts are more appropriate here. Bar charts help us see how the data is spread across different categories, making it easier to spot patterns, unusual values, or trends.

A graph of a number of people

AI-generated content may be incorrect.

A graph showing different colored bars

AI-generated content may be incorrect.

A graph of a customer service

AI-generated content may be incorrect.

A chart with different colored squares

AI-generated content may be incorrect.

A graph showing different colored squares

AI-generated content may be incorrect.

A graph of a customer service

AI-generated content may be incorrect.

The graphs above illustrate the distribution of various categorical variables. Overall, the distributions appear as expected, with no noticeable skewness or significant outliers, which is a positive sign.

A screenshot of a computer

AI-generated content may be incorrect.The data indicates that most customers are female, have high incomes, and are either married or widowed. However, the distribution of resolution status may be cause for concern. There appears to be an unusually high number of unresolved issues, which could be a contributing factor to customer churn. This potential relationship warrants further investigation to uncover any meaningful patterns.

The chi-squared test is like the Pearson correlation matrix shown earlier, however it shows relationships between categorical variables instead of numerical.

A P value of over 0.05 shows that there is a correlation between features. The results from this test are more promising, with several categories displaying strong connections, particularly with ‘ProductCategory.’ This highlights some areas of interest to be further explored in later steps.

**3 Data Cleaning & Preprocessing**

Several steps were taken to prepare the data for machine learning:

**Data Aggregation:**  
Raw transactional, service, and online activity data were aggregated at the customer level. This transformation simplified the dataset and retained key behavioural indicators, improving both modelling efficiency and interpretability.

**Outlier Flagging:**  
High-spending customers were identified using the 99th percentile threshold and flagged instead of removed or modified. This approach preserved valuable behavioural signals from distinct segments while preventing extreme values from skewing model training.

**Feature Engineering:**  
Several new features were created, including interaction recency, unresolved interaction rate, average spend per transaction, and days since last login. These enhancements were designed to extract deeper behavioural patterns that may correlate with churn risk.

**Dropping Unnecessary Columns:**  
Identifiers, raw lists, and timestamp fields with limited predictive power were removed to reduce noise and dimensionality, ensuring focus on meaningful input features.

**Encoding Categorical Features:**  
Categorical variables (e.g., gender, marital status, service usage) were transformed using one-hot encoding. This allowed the model to process non-numeric data while preserving category distinctions without introducing ordinal bias.

**Normalizing Numerical Variables:**  
Numerical features were scaled using Min-Max normalization to ensure uniform contribution to model learning. This step helps prevent models from overweighting features with naturally larger values.

**Final Preparation for Modelling:**  
With the completion of data aggregation, cleaning, feature engineering, encoding, and normalization, the dataset is now fully prepared for machine learning. The structured and enriched feature set captures essential aspects of customer behaviour, allowing for robust modelling of churn risk. This cleaned dataset provides a solid foundation for training, validating, and interpreting predictive models in the next phase of the project.

**4 Machine Learning Model Building**

**Model Selection:**

For the churn prediction task, several machine learning algorithms were considered. However, after evaluating the trade-offs between predictive performance and interpretability XGBoost (Extreme Gradient Boosting) was selected as the most suitable model due to its balance of high predictive accuracy, interpretability, resistance to overfitting, and its high imbalance handling.

**Training Process:**

The final dataset was cleaned, encoded, and split into training and test sets using an 80/20 stratified split to preserve class proportions. XGBoost was then trained with manually tuned hyperparameters, to optimise its performance and help to deal with the dataset’s class imbalance.

**Evaluation Metrics**

The model achieved strong **recall performance** on the test set, correctly identifying **95% of churners**, which is critical in churn prediction scenarios where the cost of missing a potential churn is high. This positions the model as a **reliable early-warning system** that flags at-risk customers for proactive retention efforts.

Additionally, the model demonstrated **consistent generalisation** during cross-validation, with an average F1 score of **0.31**, suggesting it maintained reasonable performance across different subsets of the data. The high recall also highlights the model's **ability to detect rare events** in an imbalanced dataset, a common challenge in churn prediction.

**5 Recommendations & Business Applications**

**Business Application:**

While the current version of the model prioritizes sensitivity, this trade-off is **strategically aligned** with business objectives: catching churners early is more valuable than risking a few false positives. The model can be integrated into the business to:

* **Score and prioritize** customers based on churn risk.
* **Trigger targeted interventions** such as loyalty offers, personalized emails, or account reviews.
* Inform marketing teams on **where to focus retention resources**.

**Model Improvements:**

The current model lays a **strong foundation** for churn prediction. Several enhancements can further boost its effectiveness:

* **Threshold Tuning**: Calibrating the decision threshold to balance recall and precision based on business tolerance for false positives.
* **Resampling Techniques**: Implementing SMOTE or other methods to address class imbalance and improve prediction confidence.
* **Explainability Tools**: Leveraging SHAP to uncover which customer behaviours most influence churn predictions — valuable for both model trust and actionable insights.
* **Feature Enrichment**: Introducing behavioural features such as session frequency, change in spending patterns, or service usage over time.

**7 Conclusion**

The developed XGBoost model demonstrates a **promising ability to identify high-risk customers** with strong recall, making it a valuable asset for churn mitigation strategies. Although there is room to improve its precision, the model’s current performance enables the business to take proactive steps in customer retention. With continued iteration and refinement, this model can become a cornerstone of the company’s customer success strategy.