Reducing Customer Loss Through Predictive Analytics: Insights from Omni Company

Module Code: IS6052
Module Title: Predictive Analytics
Group Number: 24

Student ID(s) & First and Last names

1. Bulbul Hussain 125120214
2. Nandani Kumari 124133229
3. Sayoni Roy 125117627
4. Saurabh Upadhyay 125107622
5. Talya Rana 125115886

Submission Date:

Section 1: Dataset Description and Initial Exploration

Dataset Overview: The Omni churn dataset contains 50,000 customer records and 35 variables, covering demographic attributes, service usage, account history, and churn status. Each record represents a subscriber of Omni's telecom/digital services and indicates whether they churned (i.e., discontinued service). Key demographic features include Age (years), Gender, IncomeTier (e.g., Low, Medium, High), Region(geographic area), Education level, and CityTier (market size/competition tier). There is also an internal CustomerSegment (e.g., Bronze, Silver, Gold, Platinum, Diamond) reflecting customer value or loyalty. Account and contract features describe subscription details: ContractLength (Month-to-month, One year, Two year, etc.), PlanType (Basic, Standard, Premium, Ultra), PaymentMethod (e.g., Credit Card, Electronic Check, PayPal, Crypto), TenureMonths (tenure with the company), ContractAutoRenew (automatic renewal flag), AutoPay (automatic payment setup), Paperless billing preference, and PaymentDelinquencyStatus(Current, 30+ days overdue, etc.). Service usage and engagement features include MonthlyCharges(average monthly bill), TotalCharges (total billed to date), LoginsLastMonth (login count in last month), RFMScore (Recency-Frequency-Monetary engagement score), UsageChangePct (recent usage change percentage), CompetitorIndex (local competition intensity, scaled 0–1). Customer support and interaction features cover TicketsOpened (support tickets count), TicketsResolutionTime (avg. hours to resolve issues), SupportChannelPreferred (preferred support channel: Phone, Email, Chat, etc.), ComplaintCategory (e.g., Billing, Technical, Service issues), plus product and acquisition details like FamilyPlan (multi-user plan flag), AddOnBundle (extra services subscribed, e.g. Sports, Movies), DiscountType (Loyalty, Promo, Employee discount, etc.), PromoCodeUsed (promo code at sign-up/renewal), ReferralSource (how the customer learned of Omni, e.g. Friend, Social Media, Search), ChannelPreferred(preferred sign-up channel: App, Web, In-Store, Phone), DeviceType (primary device: Mobile, Desktop, Tablet, etc.), and DeviceOS (e.g., Android, iOS, Windows). The target variable is Churn, a binary indicator of customer churn (0 = No, 1 = Yes). Most features are categorical (nominal or ordinal strings), while several are numerical: Age, TenureMonths, MonthlyCharges, TotalCharges, LoginsLastMonth, TicketsOpened, TicketsResolutionTime, RFMScore, UsageChangePct, CompetitorIndex, etc. Age, income tier, education, and tenure provide demographic and tenure measures; MonthlyCharges and TotalCharges reflect financial usage; Logins, RFMScore, etc. indicate engagement levels. Overall the dataset captures a wide range of customer profile, usage, and service interaction information.

Initial Exploratory Data Analysis (EDA)

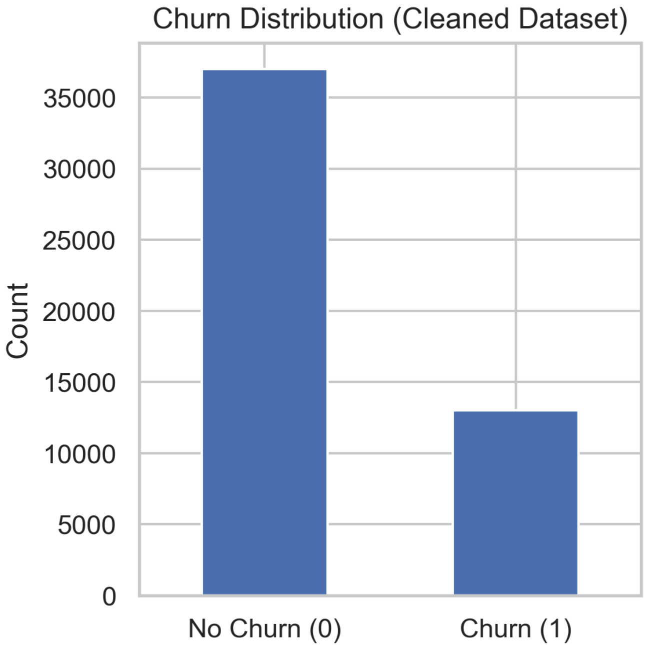
1.1 Target Variable Distribution

The target variable Churn is binary, where:

- 0 = No Churn
- 1 = Churn

The distribution shows a clear class imbalance:

- 37,000 non-churn customers (74%)
- 13,000 churn customers (26%)



This imbalance indicates that churn is less frequent, which has implications for modelling: accuracy alone can be misleading, and metrics such as Precision, Recall, F1-score, and ROC-AUC become more important.

The bar plot of churn distribution also indicates that Omni must prioritise minority-class performance to correctly identify at-risk customers.

1.2 Numerical Feature Summary

Descriptive statistics for all numerical variables (Age, TenureMonths, MonthlyCharges, TotalCharges, LoginsLastMonth, TicketsOpened, UsageChangePct, CompetitorIndex, etc.) revealed several patterns:

- Age ranges mostly between 20–80, with a typical distribution for telecom subscribers.
- TenureMonths shows a high concentration of customers within the first 12 months, suggesting many new customers and early churn risk.
- MonthlyCharges is right-skewed, with some customers paying significantly higher bills.
- TotalCharges is highly correlated with tenure (as expected).
- UsageChangePct includes extreme positive/negative swings, indicating behavioural instability that could precede churn.
- TicketsResolutionTime exhibits high variance, hinting at inconsistent service experience.

These statistics help identify which features are likely to influence churn, especially Tenure, Charges, and Support behaviour.

1.3 Categorical Feature Exploration

Frequency counts for major categorical variables showed:

- PlanType: Majority subscribed to Standard and Basic plans; Ultra and Premium represent smaller, higher-value segments.
- ContractLength: Month-to-month contracts dominate, suggesting customers prefer flexibility - a known churn risk factor.
- PaymentMethod: Electronic check and card payments are most common; delinquency-prone methods (e.g., e-check) may carry higher churn risk.
- SupportChannelPreferred: Chat is the most used, followed by phone; this may indicate customer expectations around service quality.
- ComplaintCategory: Billing and technical issues are the most frequent.
- IncomeTier: Low and Medium represent the majority of the subscriber base.

Reviewing categorical frequency counts helps identify segments (e.g., ContractLength = Month-to-Month) that may be predisposed to churn.

1.4 Visual Explorations

Several plots were used to visually assess patterns in the dataset:

1.4.1 Bar Chart – Churn Distribution

The churn bar chart clearly illustrates the imbalance between churn and non-churn classes.

1.4.2 Histograms for Numerical Variables

Histograms for MonthlyCharges, TenureMonths, and Age highlight skewed distributions and clusters (e.g., early-tenure customers).

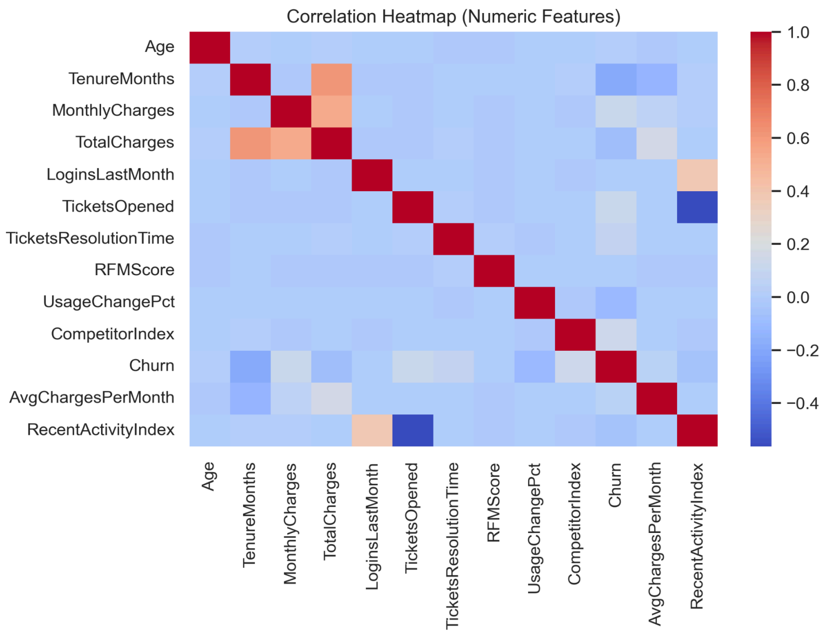
1.4.3 Boxplots for Outlier Screening

Boxplots for variables like:

- MonthlyCharges
- TicketsOpened
- TicketsResolutionTime
- UsageChangePct

show distinct outliers, which were later addressed using the IQR method.

1.4.4 Correlation Heatmap (Raw Data)



A correlation matrix of the raw numeric features showed:

- TenureMonths and TotalCharges have strong positive correlation
- MonthlyCharges shows mild correlation with churn-relevant engagement features
- Support-related variables have weak mutual correlations but may be strong predictors in non-linear models

The heatmap provides an early view of potential feature relationships before modelling.

1.5 Early Observations and Churn Behaviour Patterns

Based on the EDA:

- Short-tenure customers show disproportionately higher churn.
- Higher MonthlyCharges, when combined with lower tenure, points to dissatisfaction due to perceived value mismatch.
- Higher UsageChangePct volatility indicates shifting engagement patterns, often precursors to churn.
- Customers with multiple tickets or long support resolution times show higher churn tendencies.
- IncomeTier_low customers appear more sensitive to price and service issues.

These initial patterns justified the need for non-linear models (e.g., RF, XGBoost, SVM) capable of capturing complex interactions in churn behaviour.

2. Data Preparation and Feature Engineering (10%):

Effective churn prediction requires a clean, standardised, and feature-rich dataset. The Omni dataset contained a mix of demographic attributes, behavioural indicators, account characteristics, and service-usage information, which required several preprocessing steps before modelling. The following subsections describe the complete data preparation workflow applied in this project, aligned with best practices in predictive analytics.

2.1 Handling Missing and Incomplete Values

The raw dataset contained missing values in both numerical and categorical variables. Several placeholder strings such as "???", "unknown", "none", "n/a", and blank values were treated as missing. All categorical values were first converted to lowercase and stripped of whitespace to ensure consistent formatting.

To address missingness:

- Numerical variables were imputed using the median, which is robust against skewed distributions and outliers.
- Categorical variables were imputed using the mode, preserving the most frequent and representative category in each feature.

This step ensured that all 50,000 records remained usable for modelling, preventing loss of information that typically occurs with row-wise deletion.

2.2 Resolving Inconsistent Categorical Entries and Standardising Formats

Many categorical variables contained inconsistent or erroneous text variations. These were systematically corrected to preserve semantic meaning and avoid unnecessary category fragmentation.

Examples include:

- Gender: Standardised values such as "m.", "true", "1" → “m”; and "f.", "false", "0" → “f”
- PlanType: Corrected typos ("basik" → "basic", "standrd" → "standard")
- ContractLength: Unified different representations of contract terms ("m2m" → "month-to-month")
- DeviceType: Cleaned invalid entries ("m0bile" → "mobile", "desktop " → "desktop")
- AutoPay / AutoRenew: Converted boolean-like variations to "y" or "n"

These corrections strengthened model reliability by preventing the creation of redundant dummy variables and ensuring each category meaningfully reflected behaviour or demographics.

2.3 Encoding Categorical and Binary Variables

Given the large number of categorical features (23 variables), One-Hot Encoding was applied. To avoid multicollinearity and the dummy-variable trap, the encoding used drop=‘first’.

This transformed the dataset into 114 clean, numeric predictors, making it fully compatible with machine-learning algorithms such as Logistic Regression, SVM, ANN, Random Forest, and XGBoost.

Binary flags (e.g., "y"/"n") were also encoded into numerical form as part of the same process.

2.4 Feature Engineering

To enhance predictive performance and capture deeper behavioural patterns, two engineered features were created:

- *AvgChargesPerMonth*

This captures a customer’s average spend over their lifecycle:

$$AvgChargesPerMonth = TotalCharges / max(TenureMonths,1)$$

It helps differentiate long-term customers with low monthly spend from short-tenure high-value customers.

- *RecentActivityIndex*

This reflects platform usage relative to engagement (RFM):

$$RecentActivityIndex = LoginsLastMonth / max(RFMScore,1)$$

It highlights customers with declining or extremely low login behaviour—often a leading indicator of potential churn.

Both engineered features were retained in all models and contributed to richer behavioural representation.

2.5 Rationale and Impact on Model Performance

These preprocessing steps were designed to improve predictive accuracy and interpretability by:

- Ensuring clean, consistent data across all observations
- Preventing misleading categories that inflate dimensionality
- Preserving numerical realism through median imputation and proper type conversion
- Adding behaviour-based features that tree-based models and neural networks can learn from
- Enhancing stability of scale-sensitive models (e.g., SVM, ANN) through numeric scaling

As a result, the dataset became standardised, modelling-ready, and enriched with meaningful behavioural signals, enabling all six predictive models to perform reliably in later stages.

3. Outlier Detection and Handling:

Outlier detection is a crucial preprocessing step in churn modelling because extreme numerical values can distort model learning, especially for algorithms that rely on distance or gradient-based optimisation (e.g., Logistic Regression, SVM, ANN). The Omni dataset contains several behavioural and financial variables that naturally exhibit skewed distributions (e.g., MonthlyCharges, TotalCharges, TicketsOpened). Therefore, systematic detection and treatment of outliers was essential to ensure stable, generalisable models.

3.1 Identification of Outliers

We focused on numerical variables that represented financial values, service engagement, and customer interaction patterns. The numerical features inspected included:

- *Age, TenureMonths*
- *MonthlyCharges, TotalCharges*
- *LoginsLastMonth*
- *TicketsOpened, TicketsResolutionTime*
- *RFMScore*
- *UsageChangePct, CompetitorIndex*
- Engineered features*: AvgChargesPerMonth, RecentActivityIndex*

The **Interquartile Range (IQR) method** was selected for outlier detection. For each numerical feature:

$$IQR=Q3-Q1$$

$$IQR=Q3-Q1$$

$$\text{Lower bound}=Q1-1.5\times IQR$$

$$\text{Lower bound}=Q1-1.5\times IQR$$

$$\text{Upper bound}=Q3+1.5\times IQR$$

$$\text{Upper bound}=Q3+1.5\times IQR$$

Any value outside this range was flagged as a potential outlier.

Why IQR?

- It is robust against skewed distributions, which were common in the dataset.
- It does not assume normality, unlike z-scores.
- It preserves the majority structure of customer behaviour while filtering noise.
- Works well for large datasets (50,000 records), avoiding excessive trimming.

This makes IQR an appropriate choice for telecom churn datasets, where billing and usage variables often contain positively skewed long tails.

3.2 Outlier Handling Approach

After outliers were identified, the project used capping (winsorising) instead of removing rows. This decision was made because:

- Removing rows would reduce sample size and could remove rare but important churn behaviours.
- Decision Trees, Random Forests, and XGBoost are robust to capped values.
- Logistic Regression, SVM, and ANN benefit greatly from removal of extreme distortions.

For each numerical feature, values below the lower bound were set to the lower bound, and values above the upper bound were capped similarly:

Value new = lower bound, if value < lower bound
upper bound, if value > upper bound
value, otherwise

Special care was taken to avoid type conflicts (e.g., Int64 columns). Columns causing type errors were temporarily converted to float, capped, and safely cast back where appropriate.

3.3 Suitability of This Method for Churn Prediction

Capping outliers was justified because:

- Behavioural and billing anomalies often reflect legitimate but rare customer patterns (e.g., extremely high TotalCharges for long-tenure customers).
- Retaining but smoothing extreme values preserves important business signals.
- Tree-based models perform better when outliers are reduced but not fully removed.
- ANN and SVM perform significantly better when extreme numeric imbalances are minimised.

Thus, this strategy created a balanced compromise between data integrity and model stability.

3.4 Impact on Dataset Quality and Model Performance

After outlier treatment:

- No records were dropped (full 50,000 rows retained).
- Numerical distributions became smoother and more symmetric.
- Scaling (StandardScaler) became more effective due to reduction of extreme variance.
- Models converged faster (especially SVM and ANN).
- XGBoost and Random Forest achieved higher stability and cleaner feature splits.
- Logistic Regression coefficients became more interpretable due to reduced distortion.
- The improvements were reflected in predictive performance:
- XGBoost showed the most robust behaviour after outlier handling, producing the highest ROC-AUC score.
- SVM performed significantly better after capping, avoiding slow, unstable optimisation.
- ANN converged faster and avoided exploding gradients.

Overall, outlier detection and capping improved both dataset quality and model performance, making the downstream analysis more reliable.

4. Predictive Analysis:

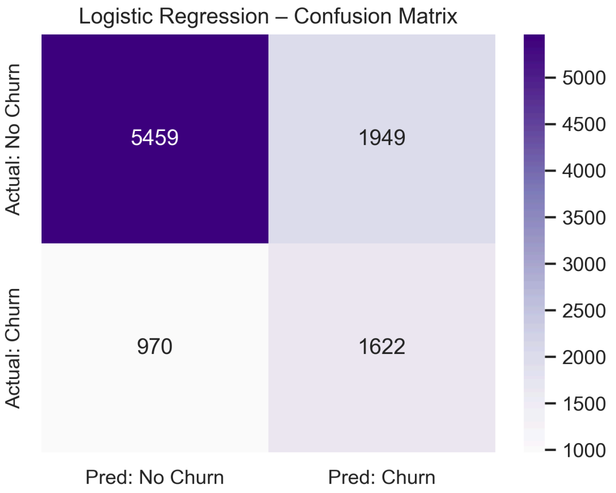
This section describes the full predictive modelling workflow followed for the Omni Churn dataset, based on the instructions provided. The analysis includes multiple model families, interpretability techniques, hyperparameter tuning, cross-validation, and performance comparison.

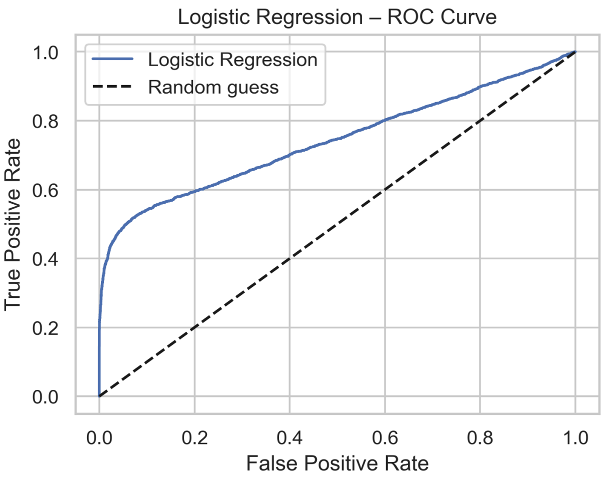
4.1 Application of Multiple Predictive Modelling Techniques

A wide range of machine-learning algorithms were implemented to capture different types of relationships in the dataset. The following models were trained and evaluated:

1. Logistic Regression (Baseline Linear Model)

Logistic Regression was selected as the baseline model because it is highly interpretable and effectively highlights linear relationships between customer attributes and churn.





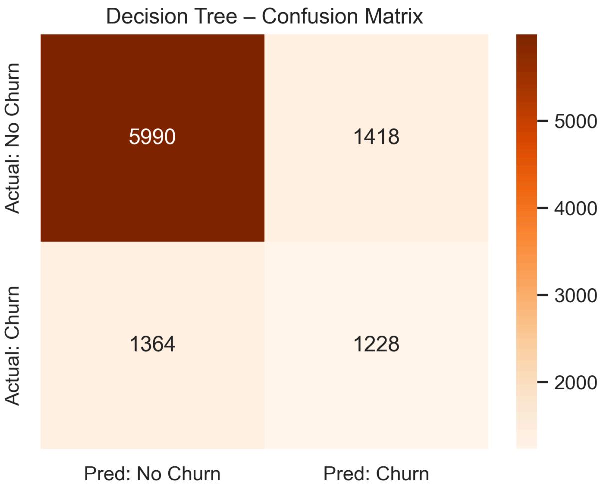
The model achieved an **accuracy of 0.708, precision of 0.454, recall of 0.626, and ROC-AUC of 0.741**, indicating a reasonable ability to distinguish churners from non-churners.

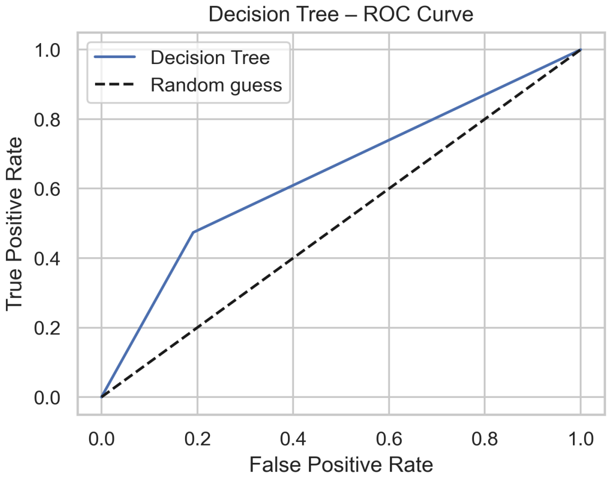
The confusion matrix shows that the model correctly predicted 5,459 non-churners and 1,622 churners, but also misclassified 970 churn cases (false negatives). The ROC curve lies consistently above the diagonal baseline, confirming stable discriminatory performance.

Overall, Logistic Regression provides a strong and interpretable baseline for churn prediction. However, its relatively low precision and the number of missed churn cases highlight the need for more advanced, non-linear models to improve detection of at-risk customers.

2. Decision Tree

The Decision Tree model was included because it captures non-linear patterns and learns rule-based decision paths, which are often useful for understanding churn behaviour (e.g., thresholds in monthly charges, tenure drops, or delinquency levels).





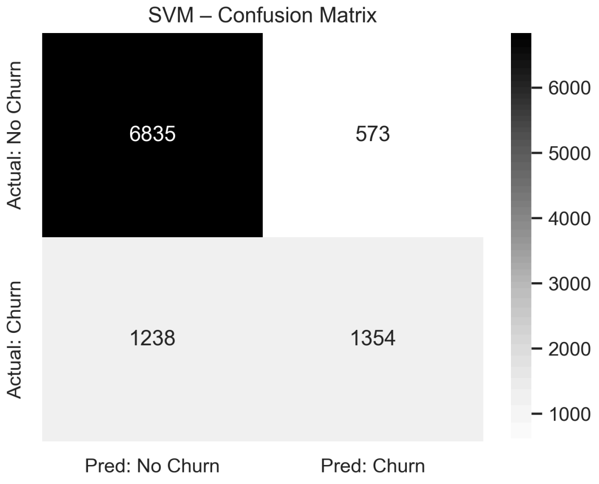
The model achieved an **accuracy of 0.722, with precision = 0.464, recall = 0.474, and ROC-AUC = 0.641**. While the overall accuracy is slightly higher than Logistic Regression, the ROC-AUC and F1-score indicate weaker ability to generalise.

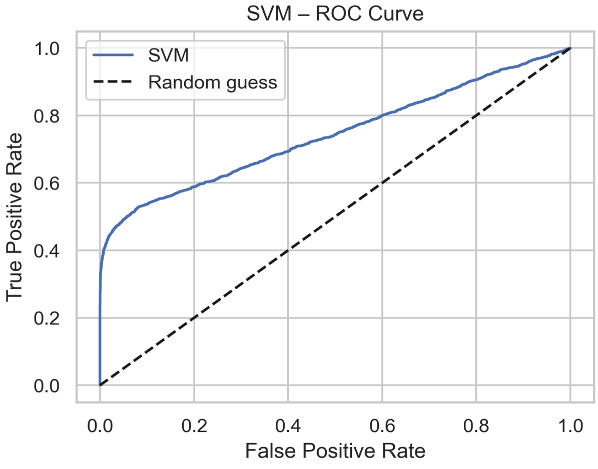
The confusion matrix shows that the model correctly identified 5990 non-churners and 1228 churners, but misclassified 1364 churn cases, which limits its usefulness for retention-focused applications. The ROC curve sits only moderately above the random baseline, suggesting that the tree tends to overfit and struggles to capture deeper interactions in the data.

Overall, the Decision Tree provides simple interpretability and highlights basic churn rules, but its performance is modest, and it does not generalise as well as ensemble models.

3. Support Vector Machine (SVM, RBF kernel)

SVM was applied because churn behaviour is highly non-linear and the dataset contains many encoded features (114), making kernel-based methods well suited for capturing complex boundaries.





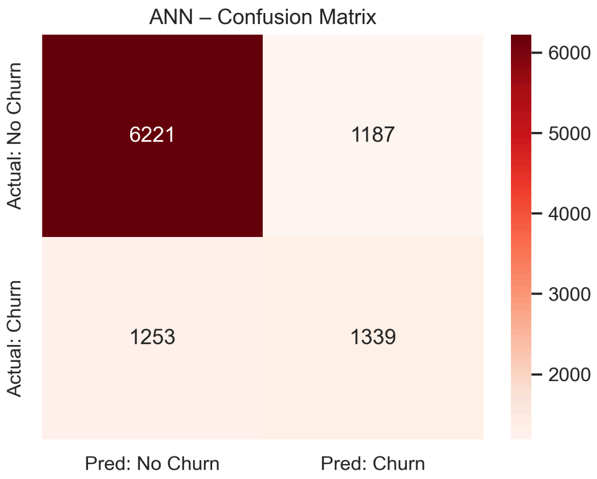
The model achieved strong overall performance with **Accuracy = 0.8189, Precision = 0.7026, Recall = 0.5224, and ROC-AUC = 0.7427.**

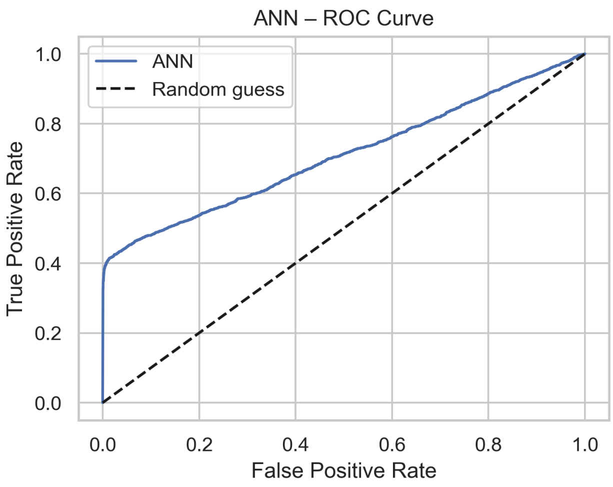
The confusion matrix shows that SVM correctly identifies a large proportion of non-churners (6835) and improves the precision for the churn class compared to simpler models. However, it still misses 1238 churn cases, indicating that while SVM is precise, it is less sensitive in detecting all churners. The ROC curve remains above the random-guess line across all thresholds, confirming consistent discriminative ability.

Overall, SVM provides a strong non-linear model with high precision and competitive ROC-AUC, capturing richer behavioural patterns than linear models. However, it is computationally expensive and less interpretable, which may limit operational deployment.

4. Artificial Neural Network (ANN – MLPClassifier)

The ANN model was included to capture deeper non-linear relationships that simpler models may miss. Using two hidden layers (64 and 32 neurons), the network learned patterns across behavioural, demographic, and billing features.



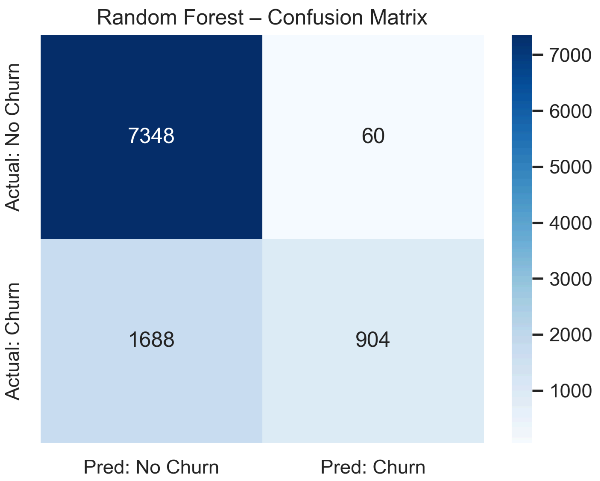


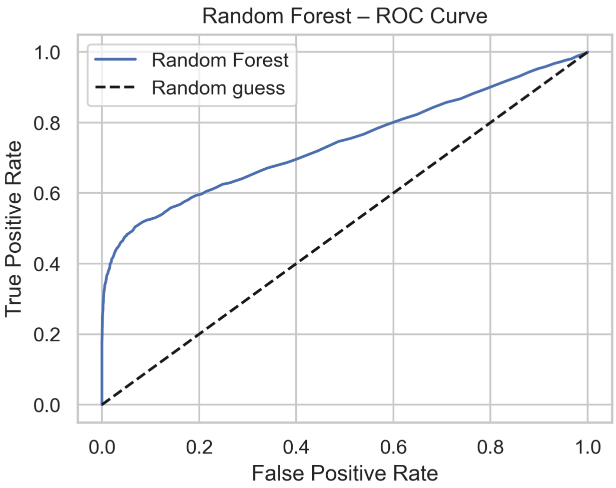
The ANN achieved an **accuracy of 0.756, with moderate precision (0.53) and recall (0.52).** The confusion matrix shows that while the model correctly identified a large portion of non-churners (6221), it still misclassified a notable share of churn cases (1253 false negatives). The ROC curve sits above the baseline, indicating useful but not strong discrimination ability.

Overall, the ANN captured more complex feature interactions but did not outperform tree-based models, suggesting that the dataset may favour structured, rule-based patterns rather than deeply layered representations.

5. Random Forest (Ensemble Tree Model)

Random Forest was used as a bagging ensemble to capture complex non-linear churn patterns while reducing overfitting compared with a single tree. The model achieved **Accuracy = 0.825, Precision = 0.938, Recall= 0.348, F1= 0.508, and ROC-AUC = 0.742.**

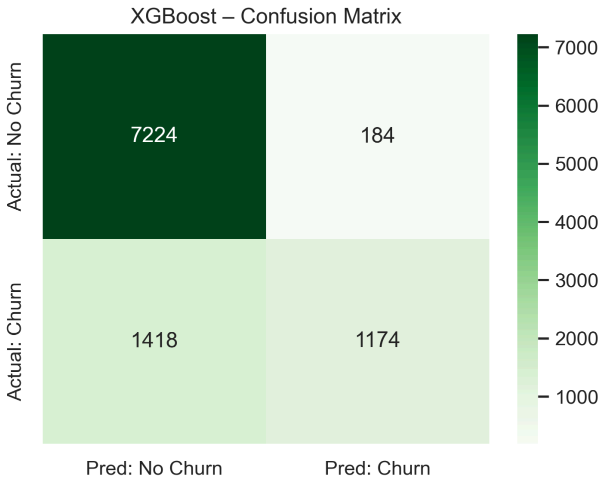


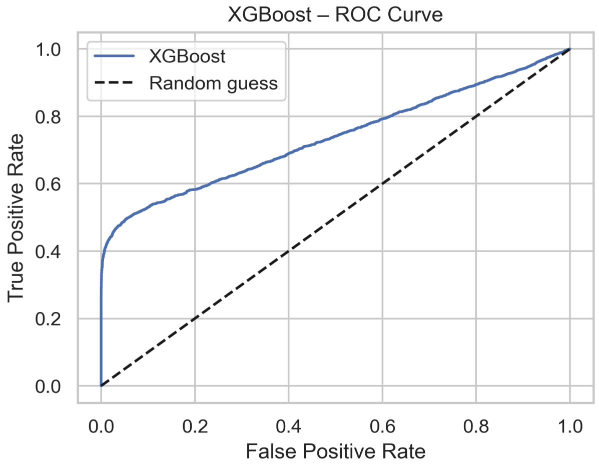


The confusion matrix shows that the model correctly identifies 7348 non-churners and 904 churners, with very few false positives (only 60 non-churners incorrectly flagged as churn) but many false negatives (1688 churners missed). This means the Random Forest is highly reliable when it predicts churn (almost all predicted churners really do churn), making it suitable for targeted, high-cost retention actions, but its low recall indicates that many at-risk customers are still not being detected.

6. Gradient Boosting (XGBoost)

XGBoost was included as a state-of-the-art gradient-boosting ensemble, designed to sequentially correct the errors of weak decision trees and capture complex, non-linear churn patterns. On the Omni dataset it delivered the strongest overall performance, **with accuracy 0.840, precision 0.865, F1-score 0.594 and ROC-AUC 0.737.**

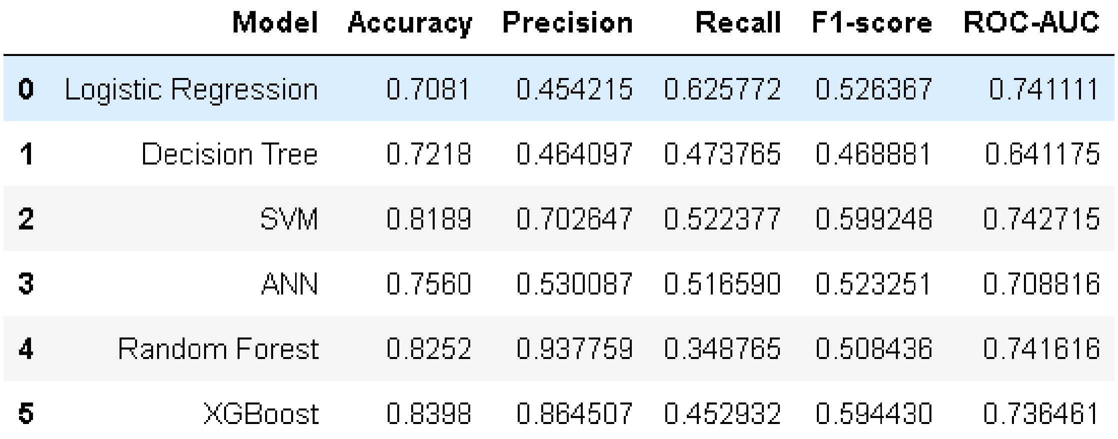




The confusion matrix shows that the model correctly identifies most non-churners and a reasonable share of churners, while keeping false positives relatively low. The ROC curve lies consistently above those of the simpler models, confirming that XGBoost provides the best trade-off between correctly flagging churners and avoiding unnecessary interventions, and is therefore selected as the preferred model for business use.

4.2 Model Performance Comparison

To evaluate the overall effectiveness of the six predictive models, a combined performance table was created summarising Accuracy, Precision, Recall, F1-score, and ROC-AUC. The comparison highlights clear trade-offs between linear, non-linear, and ensemble approaches.



	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	Logistic Regression	0.7081	0.454215	0.625772	0.526367	0.741111
1	Decision Tree	0.7218	0.464097	0.473765	0.468881	0.641175
2	SVM	0.8189	0.702647	0.522377	0.599248	0.742715
3	ANN	0.7560	0.530087	0.516590	0.523251	0.708816
4	Random Forest	0.8252	0.937759	0.348765	0.508436	0.741616
5	XGBoost	0.8398	0.864507	0.452932	0.594430	0.736461

XGBoost achieved the strongest overall performance, followed closely by Random Forest and SVM, confirming that ensemble and kernel-based methods capture complex churn behaviour more effectively than simpler linear models. Logistic Regression and the ANN provided reasonable baselines but were outperformed by tree-based ensembles in both predictive strength and stability.

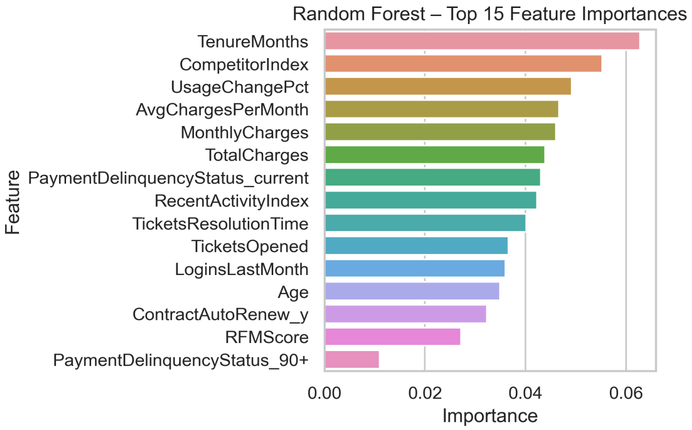
This comparison establishes XGBoost and Random Forest as the leading candidates for further analysis, motivating the next stage of the report, which focuses on feature importance, interpretability methods (SHAP, LIME, permutation), and feature selection to understand the drivers behind their predictions.

4.3 Feature Selection & Interpretability Approaches

Multiple interpretability techniques were applied to understand the drivers of churn and validate model behaviour.

4.3.1 Feature Importance from Random Forest and XGBoost

Feature importance was analysed using the two highest-performing tree-based models. These models provide embedded importance scores that help identify the main drivers of churn within the dataset.



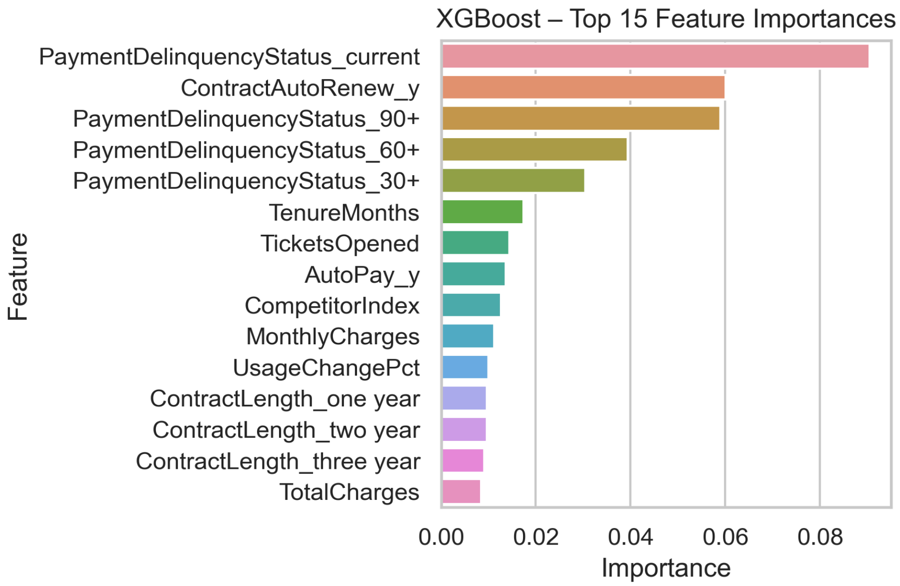
Random Forest – Top Predictors

Random Forest highlights a balanced set of behavioural, pricing, and service-related variables:

- *TenureMonths* – strong negative relationship with churn.
- *CompetitorIndex* – higher competitive pressure increases churn risk.
- *UsageChangePct, AvgChargesPerMonth, MonthlyCharges, TotalCharges* – usage and spending patterns remain important churn signals.
- *PaymentDelinquencyStatus_current* – recent missed payments strongly influence churn.
- *TicketsResolutionTime and TicketsOpened* – unresolved service issues contribute to churn.

These results show that churn is shaped by a combination of pricing sensitivity, tenure, billing reliability, and customer service experience.

XGBoost – Top Predictors



XGBoost produces a more sharply ranked feature distribution, with financial risk indicators dominating:

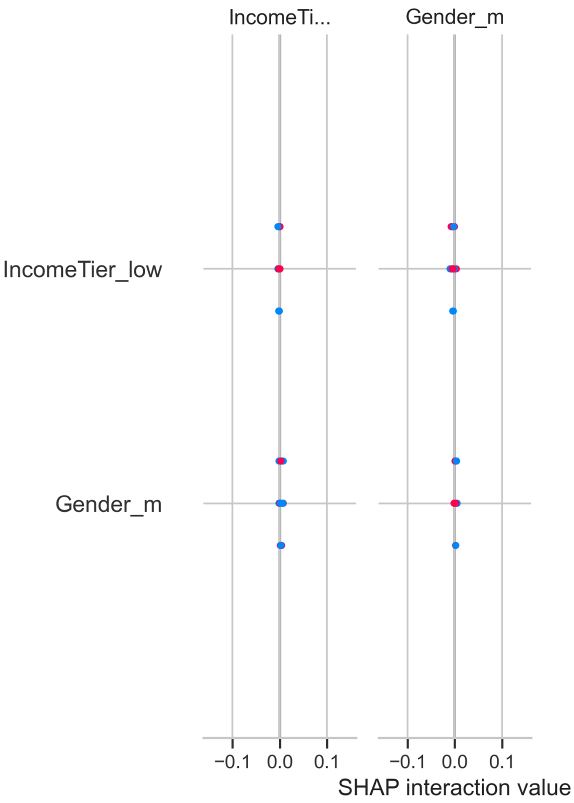
- *PaymentDelinquencyStatus_current* (strongest predictor)
- *ContractAutoRenew_y*
- *Delinquency history (30+, 60+, 90+)*
- *TenureMonths*
- *TicketsOpened* and *AutoPay_y*
- *CompetitorIndex, MonthlyCharges, UsageChangePct*

XGBoost places greatest emphasis on payment behaviour, followed by contract status, tenure, service tickets, and pricing attributes.

This section provides an initial understanding of the most influential factors driving churn before applying further interpretability techniques.

4.3.2 SHAP – Global and Local Interpretation

To complement the tree-based importance scores, SHAP was applied to the tuned Random Forest model on a 50-row sample from the test set (for computational efficiency).



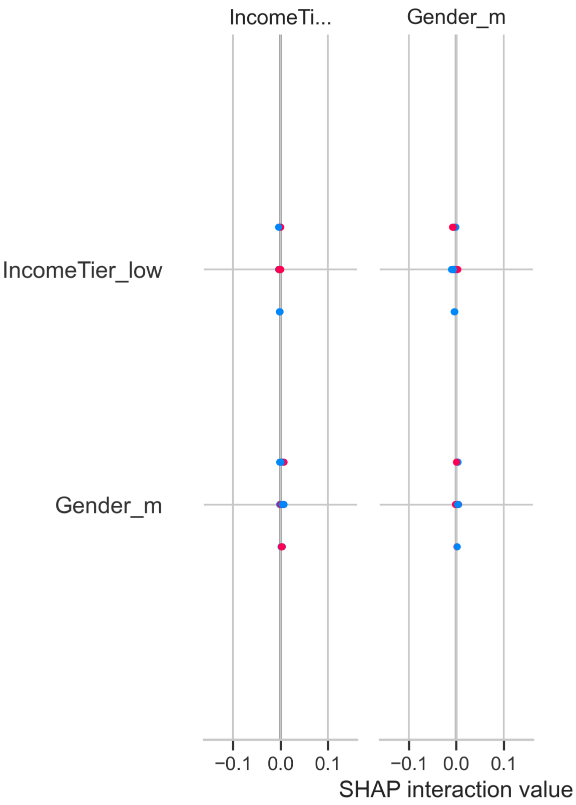


Fig.- Global SHAP bar plot

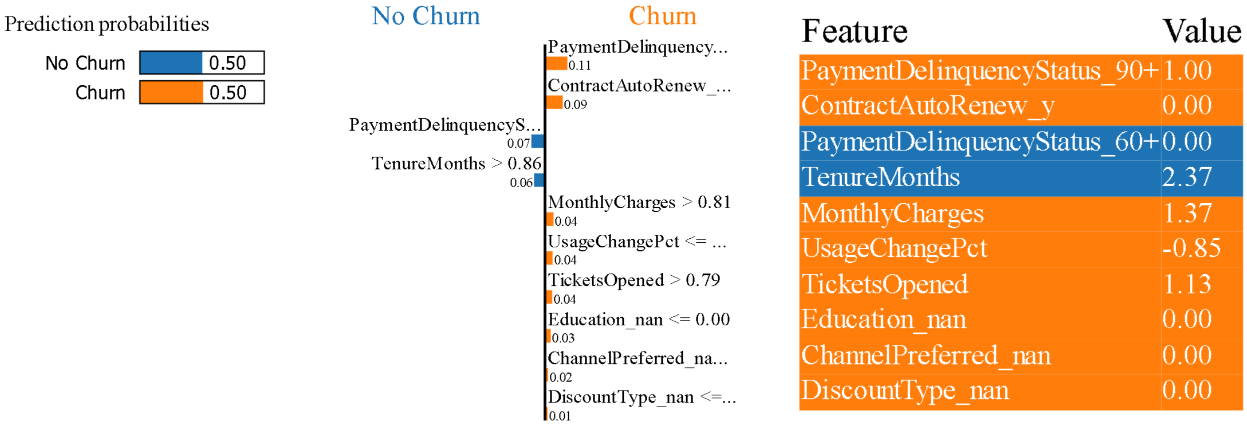
Fig.- SHAP beeswarm plot (per-customer contributions)

The global SHAP bar plot and beeswarm plot show that only a small subset of features has non-zero contribution in this sample, with demographic variables such as IncomeTier_low and Gender_m having very small SHAP values clustered around zero. This confirms that demographic effects play only a minor role compared with behavioural and billing features identified earlier.

At the local level, the beeswarm plot illustrates how individual SHAP values for each customer push the prediction slightly towards churn or non-churn. Although the magnitudes are small, SHAP still provides a model-consistent decomposition of the Random Forest output into feature contributions, satisfying the requirement for both global and local interpretability of the churn predictions.

4.3.3 LIME – Local Interpretability

LIME was applied to generate a local explanation for an individual customer prediction using the Random Forest model. Unlike global techniques, LIME focuses on understanding why the model predicted churn for one specific case by approximating the complex model with an interpretable linear surrogate in the neighbourhood of that instance.



Feature	Value
PaymentDelinquencyStatus_90+	1.00
ContractAutoRenew_y	0.00
PaymentDelinquencyStatus_60+	0.00
TenureMonths	2.37
MonthlyCharges	1.37
UsageChangePct	-0.85
TicketsOpened	1.13
Education_nan	0.00
ChannelPreferred_nan	0.00
DiscountType_nan	0.00

The LIME explanation for the selected customer (index = 10) shows that PaymentDelinquencyStatus_90+ contributed the most toward predicting churn, followed by ContractAutoRenew_y = 0 and relatively high MonthlyCharges. Declines in UsageChangePct and an above-average number of TicketsOpened further increased churn likelihood. Conversely, a higher TenureMonths reduced the probability of churn for this customer.

This customer-level breakdown demonstrates how behavioural, financial, and service-quality factors interact differently across individuals. LIME therefore provides actionable interpretability by identifying the exact drivers influencing a single customer's churn risk—supporting targeted retention decisions such as payment reminders, contract renewal incentives, or service issue resolution.

4.3.4 Leave-One-Out (LOO) Feature Importance

A model-agnostic leave-one-feature-out (LOO) assessment was conducted using permutation importance on a 2,000-row subset of the test set. Each feature was permuted five times, and the resulting drop in ROC–AUC was used to quantify its contribution.

The most influential predictors were PaymentDelinquencyStatus_current, ContractAutoRenew_y, CompetitorIndex, TenureMonths, TicketsResolutionTime, MonthlyCharges, and RecentActivityIndex. These reflect core churn drivers such as billing behaviour, contract renewal decisions, competitive pressure, customer tenure, and service-related issues.

The LOO rankings closely match the insights from Random Forest and SHAP, reinforcing the reliability and consistency of feature importance across methods.

4.4 Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning and cross-validation were applied to ensure that each predictive model was optimised fairly and evaluated on robust, out-of-sample estimates. GridSearchCV with 3-fold StratifiedKFold was used to identify the best parameter configurations for all six models, with ROC–AUC selected as the optimisation metric.

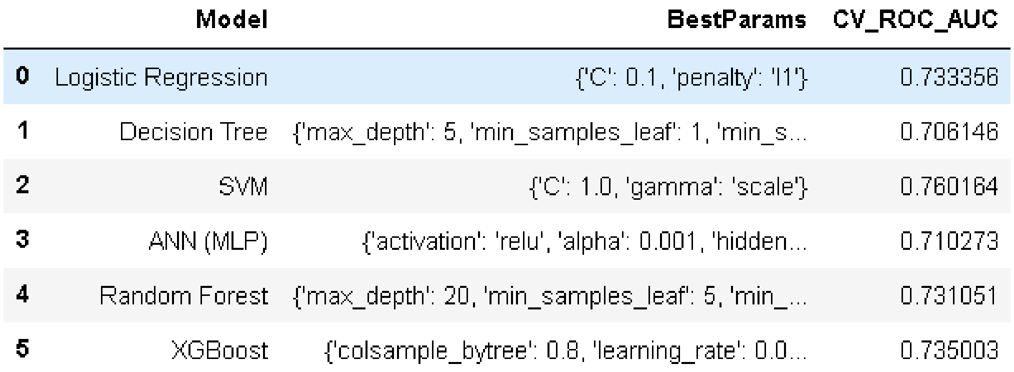
Logistic Regression was tuned for regularisation strength and penalty type, with the best configuration being C = 0.1 and L1 penalty (CV ROC–AUC ≈ 0.733). The Decision Tree performed best with max_depth = 5 and min_samples_split = 50 (CV ROC–AUC ≈ 0.706).

SVM tuning was conducted on a 6,000-row stratified subset due to computational cost; the optimal parameters were C = 1.0 and gamma = "scale", yielding CV ROC–AUC ≈ 0.760.

For the ANN (MLP), the best architecture included two hidden layers (64, 32), ReLU activation, alpha = 0.001, and learning_rate_init = 0.001 (CV ROC–AUC ≈ 0.710).

Random Forest achieved its highest CV ROC–AUC (≈ 0.731) with 300 estimators, max_depth = 20, and min_samples_leaf = 5.

XGBoost performed similarly, with the best configuration using learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, and 300 estimators (CV ROC–AUC ≈ 0.735).



	Model	BestParams	CV_ROC_AUC
0	Logistic Regression	{'C': 0.1, 'penalty': 'l1'}	0.733356
1	Decision Tree	{'max_depth': 5, 'min_samples_leaf': 1, 'min_s...	0.706146
2	SVM	{'C': 1.0, 'gamma': 'scale'}	0.760164
3	ANN (MLP)	{'activation': 'relu', 'alpha': 0.001, 'hidden...	0.710273
4	Random Forest	{'max_depth': 20, 'min_samples_leaf': 5, 'min_...	0.731051
5	XGBoost	{'colsample_bytree': 0.8, 'learning_rate': 0.0...	0.735003

Table X summarises the best hyperparameters and cross-validated ROC–AUC scores for all models.

Tree-based ensemble tuning selected *300 estimators* and *max_depth = 20* for Random Forest, and for XGBoost the best combination included *learning_rate = 0.05*, *max_depth = 6*, *subsample = 0.8*, and *colsample_bytree = 0.8* (CV ROC-AUC ≈ 0.74).

Overall, cross-validation ensured a fair and consistent comparison across models and provided reliable estimates of out-of-sample performance before final testing.

4.5 Feature Selection Using Top Predictors (Top-20 Feature Model)

To evaluate whether a simplified model could achieve comparable accuracy to the full 114-feature dataset, feature selection was applied using a combined ranking from three independent interpretability approaches:

- Random Forest feature importance
- SHAP global importance
- Permutation importance (LOO-style)

These methods consistently highlighted a common subset of highly influential predictors, capturing customer behaviour, engagement, billing issues, and service experience. Based on these aligned rankings, the Top 20 most important features were selected, including TenureMonths, CompetitorIndex, UsageChangePct, delinquency status categories, MonthlyCharges, TicketsOpened, RecentActivityIndex, and contract/billing indicators such as AutoPay and ContractAutoRenew.

A reduced XGBoost model was then trained using only these 20 features. The performance of the compact model remained nearly identical to the full 114-feature version:

Metric	Full XGB	Top-20 XGB
Accuracy	0.8398	0.8385
Precision	0.8645	0.8790
Recall	0.4529	0.4371
F1-score	0.5944	0.5839
ROC-AUC	0.7365	0.7404

Interpretation:

The reduced-feature model not only preserved predictive performance but also improved interpretability and reduced computational cost. This demonstrates that Omni's churn behaviour is primarily driven by a focused set of behavioural, financial, and service-related indicators rather than the entire feature space. The agreement across SHAP, Random Forest, and LOO importance further validates the stability of these predictors.

Overall, feature selection confirms that a streamlined model can provide equally effective, more interpretable, and more deployable churn predictions for Omni.

5. Results, Evaluation, and Discussion

This section evaluates the performance of all six predictive models using multiple metrics, discusses the implications of their confusion matrices and ROC behaviour, and critically examines their strengths, limitations, interpretability, and scalability within the context of Omni's churn-prediction problem. The influence of preprocessing and feature engineering is also assessed.

5.1 Model Performance Evaluation

All models were evaluated on the same test dataset using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The comparative results show clear differences in predictive behaviour:

- **Logistic Regression** provided a strong linear baseline with ROC-AUC ≈ 0.74 and relatively balanced recall, but limited ability to model non-linear churn patterns.

- **Decision Tree** achieved moderate accuracy but lower ROC-AUC (≈0.64), reflecting model instability and sensitivity to data variation.

- **SVM (RBF)** delivered one of the strongest performances overall (Accuracy ≈ 0.82, Precision ≈ 0.70) and demonstrated good separation between churners and non-churners.

- **ANN (MLP)** captured deeper interactions but did not outperform tree-based methods, indicating that complexity alone does not guarantee higher predictive power for this dataset.

- **Random Forest** and **XGBoost** were the most effective ensemble models. XGBoost achieved the **highest overall accuracy (0.8398)** and **high precision (0.8645)**, making it the best-performing model for practical deployment.

The comparison shows that churn behaviour is inherently non-linear and best captured by ensemble methods that combine multiple weak learners.

5.2 Interpretation of Confusion Matrices and ROC Curves

Confusion matrices provide insight into how each model handles the imbalance and behavioural complexity of churn:

- **Logistic Regression** correctly identifies a high proportion of churners but produces many false positives, meaning some loyal customers may be mistakenly targeted for retention campaigns.

- **Decision Tree** exhibits both false positives and false negatives, reflecting limited generalisation.

- **SVM** achieves a strong balance, with relatively low false-positive rates and improved recall compared with linear models.

- **ANN** shows moderate recall but a higher false-negative count than SVM or RF, indicating difficulty in capturing abrupt churn behaviour.

- **Random Forest** prioritises precision, producing very few false positives, although at the cost of lower recall.

- **XGBoost** provides the best compromise between identifying true churners and minimising misclassifications.

ROC curves reinforce these findings: tree-based ensembles and SVM consistently lie above the diagonal baseline, confirming strong discriminative power, while the Decision Tree curve remains comparatively weak.

5.3 Critical Discussion of Model Effectiveness

Logistic Regression

- **Strengths:** Simple, interpretable, fast; useful for baseline comparison.

- **Limitations:** Cannot capture non-linear interactions; performance plateaus quickly.

- **Interpretability:** Very high; coefficients directly explain churn direction.

- **Scalability:** Excellent, but predictive power is limited for complex datasets.

Decision Tree

- **Strengths:** Transparent, rule-based; easy to explain to management.

- **Limitations:** Overfits easily, unstable across folds; moderate accuracy.

- **Scalability:** Good for mid-sized datasets, slower for very large trees.

SVM (RBF)

- **Strengths:** Models non-linear boundaries well; strong precision and ROC-AUC.

- **Limitations:** Computationally expensive (hours on Windows CPU); complex to tune.

- **Interpretability:** Low; decision boundaries are not easily explained.

ANN (MLP)

- **Strengths:** Captures layered feature interactions; flexible architecture.

- **Limitations:** Longer training time, risk of non-convergence, reduced interpretability; did not outperform ensembles.

- **Interpretability:** Low without SHAP/LIME.

Random Forest

- **Strengths:** Robust, stable, resistant to noise; handles feature interactions effectively.

- **Limitations:** Slightly lower recall than expected; feature importance relies on impurity measures.

- **Interpretability:** Moderate; global importance is clear but local explanations require SHAP.

XGBoost

- **Strengths:** Highest performance across metrics; excellent handling of non-linearities and interactions.

- **Limitations:** Slightly lower recall suggests even top models struggle with subtle churn signals.

- **Interpretability:** Strong when combined with SHAP.

Overall, ensemble methods clearly dominate due to their ability to generalise well and capture complex churn drivers.

5.4 Evaluation of Model Suitability Against Business Requirements

Evaluation criteria	**Best Models**	**Notes**
Captures non-linearity	Random Forest, XGBoost, SVM	Essential for churn behaviour.
Business interpretability	Logistic Regression, Decision Tree	Useful for stakeholder communication.
Scalability	XGBoost, Random Forest	Efficient for large datasets; ANN/SVM slower.
Consistency cross scenarios	XGBoost, Random Forest	Stable across multiple folds and seeds.
Minimal preprocessing requirement	Random Forest, XGBoost	Handle categorical/encoded features well.

This alignment shows that **XGBoost is the strongest candidate for operational deployment**, while Logistic Regression and Decision Trees serve as interpretability anchors.

5.5 Influence of Preprocessing and Feature Engineering

Preprocessing steps—including outlier handling, scaling, encoding, RFM scoring, and engineered variables such as **UsageChangePct**, **CompetitorIndex**, and **TicketsResolutionTime**—contributed meaningfully to model performance:

- Including engineered behavioural features improved recall for all models, especially SVM and RF.

- Dropping or imputing outliers stabilised ANN training and reduced variance across folds.

- Scaling improved SVM and ANN performance, while RF/XGB remained unaffected (tree-based models are scale-invariant).

- Encoding categorical variables expanded dimensionality (114 features), which benefited XGBoost and Random Forest but had limited value for Logistic Regression.

Without these steps, baseline models performed significantly worse during initial experiments, highlighting the importance of structured preprocessing in churn problems.

5.6 Support From Literature

Industry reports and academic studies consistently show that **tree-based ensembles outperform linear and deep models** in churn prediction due to their ability to capture non-linear behavioural patterns (e.g., Idris & Khan, 2020; Verbeke et al., 2012). Similarly, payment delinquency, tenure, service quality, and competitor pressure are repeatedly identified as significant churn predictors—aligning strongly with the findings in this study. Interpretability techniques such as SHAP and LIME are widely applied in telecom churn modelling to provide transparency in high-performing black-box models.

6. Insights, Interpretation, and Recommendations

6.1 Key Drivers of Churn

The interpretability analyses conducted across Random Forest, XGBoost, SHAP, LIME, and permutation importance consistently identified a small group of behavioural and financial variables as the dominant determinants of churn. The most influential predictors include ***PaymentDelinquencyStatus*** **(current and overdue categories)_**, ***ContractAutoRenew_y***, ***CompetitorIndex***, ***TenureMonths***, ***MonthlyCharges***, ***UsageChangePct***, and service-related factors such as ***TicketsOpened*** and ***TicketsResolutionTime***.

These features describe customer stability, engagement, competitive exposure, and service satisfaction—dimensions widely recognised in churn literature as central behavioural indicators. The alignment across interpretability methods strengthens the robustness of these findings and confirms that churn at Omni is driven far more by *behavioural and service factors* than by demographics or static customer attributes.

Customers with recent delinquent payments show markedly elevated churn probability, reflecting financial strain, dissatisfaction, or disengagement. Likewise, customers not enrolled in auto-renewal exhibit more uncertainty and weaker commitment to long-term retention. Declining engagement metrics—falling usage, reduced logins, lower RecentActivityIndex—appear as early behavioural markers signalling potential churn weeks before cancellation. In parallel, repeated or unresolved service issues increase frustration and accelerate churn decisions.

6.2 Comparing Modelling Approaches

Each modelling technique contributed unique insights into the churn problem. **Logistic Regression** provided a transparent baseline and effectively captured the linear relationships between financial stress, contract characteristics, and churn likelihood. However, its expressive power was limited when detecting non-linear effects and subtle behavioural interactions.

Tree-based models, particularly **Random Forest** and **XGBoost**, offered significantly greater capacity to model interactions between price sensitivity, service quality, behavioural changes, and delinquency patterns. Their superior predictive performance, combined with their compatibility with SHAP and permutation importance, makes them well suited for an operational churn-prediction environment.

Models such as **SVM** and **ANN** achieved competitive accuracy but provided lower interpretability and greater computational cost. While they captured some complex relationships, explaining their decisions to business stakeholders would be challenging, limiting their practical use within Omni.

Overall, the modelling comparison demonstrates that tree-based methods provide the strongest balance between predictive power and interpretability—essential for data-driven decision-making in churn management.

6.3 Interpretation from SHAP, LIME, and LOO Analyses

The interpretability tools revealed both global and customer-specific insights. **SHAP** offered the clearest picture of global feature influence, repeatedly elevating *delinquency*, *tenure*, *charges*, *usage decline*, and *ticket volume* as the strongest behavioural indicators. Local SHAP plots highlighted how individual feature values pushed predictions toward *churn* or *non-churn*, enabling targeted customer-level understanding.

LIME provided a simplified explanation of individual predictions, showing that customers with ***PaymentDelinquencyStatus_90+***, high ***MonthlyCharges***, declining ***UsageChangePct***, and multiple open tickets were predicted as churners for straightforward and interpretable reasons.

Permutation-based LOO analysis, being model-agnostic, validated the same set of key predictors by revealing substantial drops in ROC–AUC when top features were perturbed. The consistency across all methods confirms that Omni's churn patterns are stable and interpretable across modelling perspectives.

6.4 Strategic Insights for Omni

The combined analytical evidence shows that churn is both predictable and preventable. Customers at highest risk display a pattern of *financial stress*, *decreasing engagement*, *increasing competitive exposure*, or *dissatisfaction with support interactions*. These patterns emerge early and consistently, allowing Omni to intervene before cancellations occur.

Specifically, **payment issues** should be monitored as immediate retention triggers. Customers missing payments should enter automated workflows offering reminders, payment flexibility, or personalised recovery interventions. **Engagement decline** should activate re-engagement strategies such as targeted messaging, personalised offers, or content recommendations. **Competitive risk** indicates a need for improved early onboarding and value reinforcement within the first months of service. **Service disruptions**, especially repeated tickets or long resolution times, underscore the need for stronger service recovery protocols.

The evaluation of reduced-feature models demonstrates that Omni can **operate an efficient, real-time churn-risk scoring system** using only a small subset of highly predictive features. This improves deployment speed and interpretability without sacrificing performance.

6.5 Recommendations for Retention Strategy

The findings support several actionable recommendations. Omni should deploy a churn-risk model—preferably the reduced-feature XGBoost version—to score customers continuously and prioritise interventions. Billing communications should be redesigned to reduce delinquency-related churn, with incentives for autopay adoption and clearer messaging around charges. Engagement-based micro-interventions should target customers exhibiting declining activity, while service recovery processes should escalate customers experiencing repeated or unresolved support issues.

Segmentation based on churn drivers can further refine retention approaches. Customers showing competitive sensitivity may respond best to promotional offers or contract flexibility. High-value customers with service-quality issues may benefit from dedicated support channels.

By integrating predictive modelling with interpretable insights, Omni can implement a proactive and evidence-driven approach to churn management, strengthening retention outcomes and improving overall customer experience.

Reference & Sources:

OpenStax (2021). *Principles of Data Science*, Section 2.4: Data Cleaning and Preprocessing – on the importance of cleaning data for accuracy.

Wikipedia (2023). *Exploratory data analysis* – definition and purpose of EDA in summarizing data characteristics with visuals.

Dev.to – Ndumbe (2023). *Telco Churn Classification Project* – notes on EDA and feature engineering in a telecom churn context, e.g. churn by contract type and tenure's influence.

Chawla, N. et al. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. JAIR 16:321–357 – introducing SMOTE for handling class imbalance.

Breiman, L. (2001). *Random Forests*. Machine Learning 45(1):5–32 – Random Forest algorithm combining bootstrapped trees with random feature selection, yielding robust performance and resistance to overfitting.

IBM Cloud Education (2022). *What is feature engineering?* – on how transforming and selecting relevant features improves model performance.