Swan Telecom Churn Prediction — Summary Report

Structured overview of modelling decisions, evaluation, and insights

1. Objective

Identify high-risk customers likely to churn, enabling targeted retention strategies and efficient resource allocation.

2. Modelling Pipeline

Models Considered

- Logistic Regression (baseline)
- Random Forest (final selection)

Workflow

- Train/test split for performance evaluation
- Final retraining on full dataset for scoring
- Threshold set at 0.6 to prioritise confident churn predictions

3. Model Selection Rationale

Although Logistic Regression showed slightly stronger test-set metrics (e.g. higher recall), Random Forest was selected for deployment based on the following:

- Feature interpretability: Random Forest provides direct feature importance scores, supporting stakeholder insight and strategic targeting.
- Non-linearity and interactions: It captures complex relationships between customer behaviours and churn risk that Logistic Regression may miss.
- Probability calibration: Random Forest produced a more informative churn probability distribution, with clearer separation between low and high-risk customers.
- Operational confidence: The model's ranking ability (ROC-AUC) remained strong, and its predictions aligned well with business expectations.

This decision balances statistical performance with interpretability, stakeholder trust, and deployment readiness.

4. Evaluation Metrics (Test Set)

Metric	Value	Interpretation
Accuracy	80.77%	Overall correctness of predictions
Precision (Churn)	67.58%	Most flagged churners are truly at risk
Recall (Churn)	52.94%	Half of actual churners correctly identified
F1 Score (Churn)	59.37%	Balanced view of precision and recall
ROC-AUC	84.72%	Strong ability to rank churn risk

Evaluation based on unseen test data — realistic and stakeholder-trustworthy.

5. Feature Importance — Random Forest

Top drivers of churn (visualised):

- Total Charges
- Monthly Charges
- Tenure Months
- Contract type
- Payment Method
- Online Security
- Tenure Bucket
- Internet Service
- Tech Support
- Dependents
- Partner

Supports targeted retention strategies - e.g. promote long-term contracts or bundle security services.

6. Churn Probability Distribution

Visualised across all customers

- Clear separation between low and high risk
- Annotated zones:

    - Low Risk: 0.0–0.4
    - Medium Risk: 0.4–0.6
    - High Risk: 0.6–1.0
- Threshold line at 0.6 for confident churn flagging

Strategic segmentation for campaign targeting and resource allocation.

7. Threshold Rationale (0.6 or 60%)

- Prioritises precision over recall
- Reduces false positives
- Focuses retention efforts on high-confidence churners
- Aligns with business goals: efficient, targeted outreach

8. Methodology Summary

Step-by-step process

1. Data cleaned and imputed using constant strategy
2. Train/test split applied for model evaluation
3. Logistic Regression and Random Forest trained and compared
4. Random Forest selected for deployment based on interpretability and confidence
5. Final model retrained on full dataset for scoring
6. Churn probabilities exported and visualised
7. Threshold of 0.6 applied to flag high-risk customers
8. Outputs prepared for presentation and action

Churn Modelling Phases and Data Inclusion

Phase	Data Used	Includes Churned?	Purpose
Evaluation	X_test, y_test	Yes	Compare model performance
Final Scoring	df_clean[df['Churn Value']==0]	No	Predict churn risk for active customers

This table outlines the distinction between evaluation and deployment phases, clarifying which customer segments were used and why.

9. Deliverables

- Annotated notebooks
- Visuals: ROC curve, confusion matrix, feature importance, probability distribution in one PDF document
- Churn risk scores – Top 500 (1) and All Active Customers(1) in CSV format
- Presentation-ready slide deck