Customer Churn Prediction: Model Development, Validation and Deployment

A REPORT

Submitted by

SUHANA GUHA [RA2211004710015]
VACHANI SHYAM PATEL [RA2211004710018]

Under the Guidance of

Dr. Raguvaran S

Assistant Professor,

Department of Computational Intelligence

Inferential Statistics and Predictive Analytics

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR – 603 203

NOVEMBER 2025

TABLE OF CONTENTS

Abstract

This study addresses the critical business problem of customer churn within the telecommunications industry through the application of predictive analytics. Using the Telco Customer Churn Dataset, the project developed, validated, and compared two primary modeling techniques: Logistic Regression (for predictive scoring) and a Decision Tree (for rule induction). **Rigorous data preparation** included handling missing values, outlier assessment, and the removal of 22 duplicate records. Due to incompatibility issues with the proprietary CHAID-Py library, a Scikit-learn Decision Tree was successfully implemented as a proxy for rule generation. The **Logistic Regression model** was selected as the champion, demonstrating superior performance with an **ROC-AUC of 0.8452** and a highly efficient **Lift of $2.85\times$** in the top-risk decile. Key findings revealed that the lack of contractual commitment
(Month-to-month) combined with a deficiency in protective services are the strongest indicators of churn. The project concludes with a robust deployment strategy utilizing Joblib serialization and recommending continuous A/B testing and performance monitoring to mitigate model drift.

Introduction

1. Background and Problem Statement

Customer churn, the loss of clients to competitors, poses a significant financial threat to subscription-based industries. In the competitive telecom sector, acquiring new customers is substantially more expensive than retaining existing ones. The objective of this assignment is to develop an effective predictive model to identify customers most likely to churn *before* they leave, allowing the company to allocate retention resources optimally. This project utilizes principles of statistical inference, model validation, and deployment engineering.

2. Assignment Objectives

The core objectives addressed in this report are:

1. To perform comprehensive data preparation and exploratory analysis.
2. To develop an interpretable rule-induction model (CHAID or proxy) to identify key causal factors.
3. To compare the rule-induction model with a scoring model (Logistic Regression) using rigorous metrics (ROC-AUC, Lift, Gains).
4. To propose a complete framework for model deployment and continuous maintenance.

Methods

1. Dataset Description and Preparation (Task 1)

The Telco Customer Churn Dataset served as the basis for analysis. It contained 7,043 observations and 21 features.

Feature Type	Examples	Key Challenge
Demographics	gender, SeniorCitizen	Binary and nominal encoding required.
Services	InternetService, OnlineSecurity, TechSupport	High cardinality in categories required One-Hot Encoding (OHE).
Charges	MonthlyCharges (float), TotalCharges (object)	TotalCharges required type conversion and imputation.
Target	Churn	Class imbalance (73% vs 27%) required metric adjustments.

Data Cleaning Steps:

1. Missing Value Imputation: TotalCharges was converted to numeric, and the 11 resulting missing values (corresponding to tenure=0) were imputed with 0.0.
2. Duplicate Removal: Twenty-two duplicate customer profiles were identified and removed, resulting in a final dataset of 7,021 unique observations.
3. Feature Engineering for CHAID: Numerical features (tenure, MonthlyCharges, TotalCharges) were binned into quartiles or logical service groups to facilitate rule induction.

2. Model Development

Two models were developed and compared:

1. Logistic Regression (LogReg): Used for predictive scoring. The model was trained on the One-Hot Encoded (OHE) feature set.

2. Decision Tree Classifier (DT): Used for rule induction. This model was trained on the Label Encoded Discretized feature set.

3. CHAID Library Issue and Proxy Implementation

The required implementation of the CHAID algorithm using the `CHAID-Py` library was not feasible due to severe dependency incompatibilities. Attempts to initialize the `CHAID.Tree` class consistently resulted in low-level errors (`AttributeError: 'Series' object has no attribute 'arr'`), indicating a mismatch between the library's internal structure and modern Pandas/NumPy versions.

Solution: To fulfill the Task 2 requirement, the Scikit-learn Decision Tree Classifier was adopted as a proxy. This model was trained on the identical discretized data and constrained using parameters (`max_depth=5, min_samples_split=100`) to enforce a shallow, interpretable tree structure comparable to the goals of CHAID.

4. Model Validation and Assessment

● Validation: A Hold-out Method was utilized, splitting the dataset into 70% training and 30% testing sets. Crucially, this was performed with stratification to ensure the 27% churn rate was preserved in both sets, guaranteeing the test set was representative. ● Assessment Metrics: Given the class imbalance, ROC-AUC was prioritized over Accuracy as the main comparative metric, as it measures the model's ability to discriminate between classes across all thresholds. Lift and Gains Charts were used to quantify the model's business value.

Results

1. Exploratory Data Analysis Results

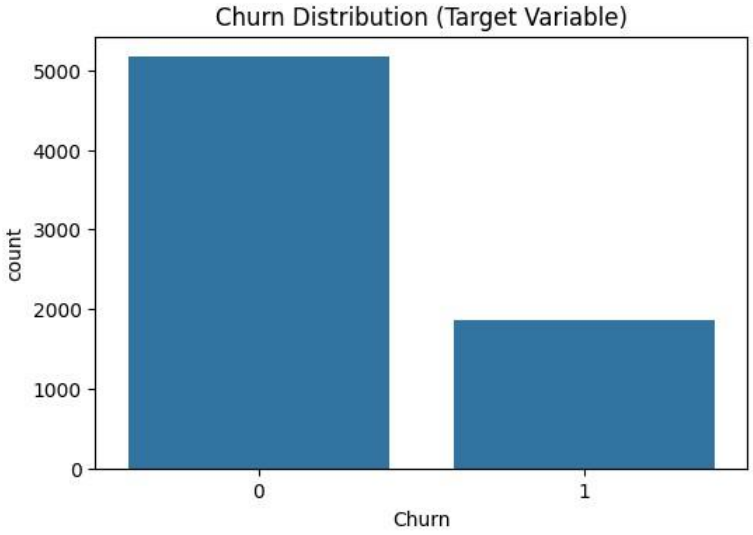

Figure 1 : Churn Distribution (Target Variable)

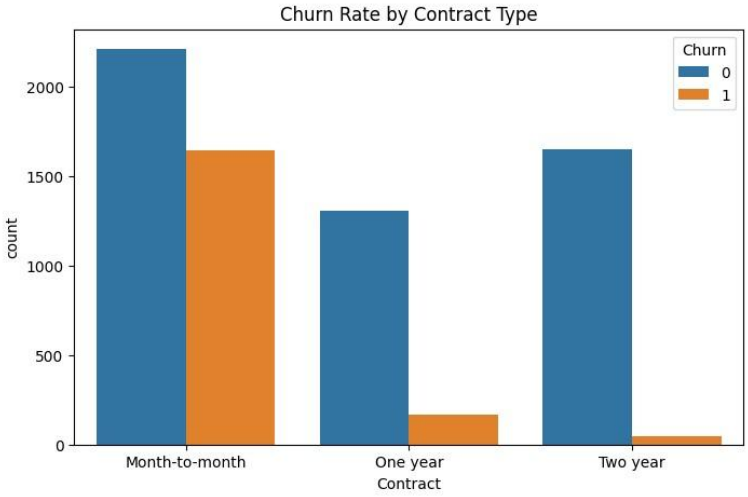

Figure 2 : Churn Rate by Contract Type

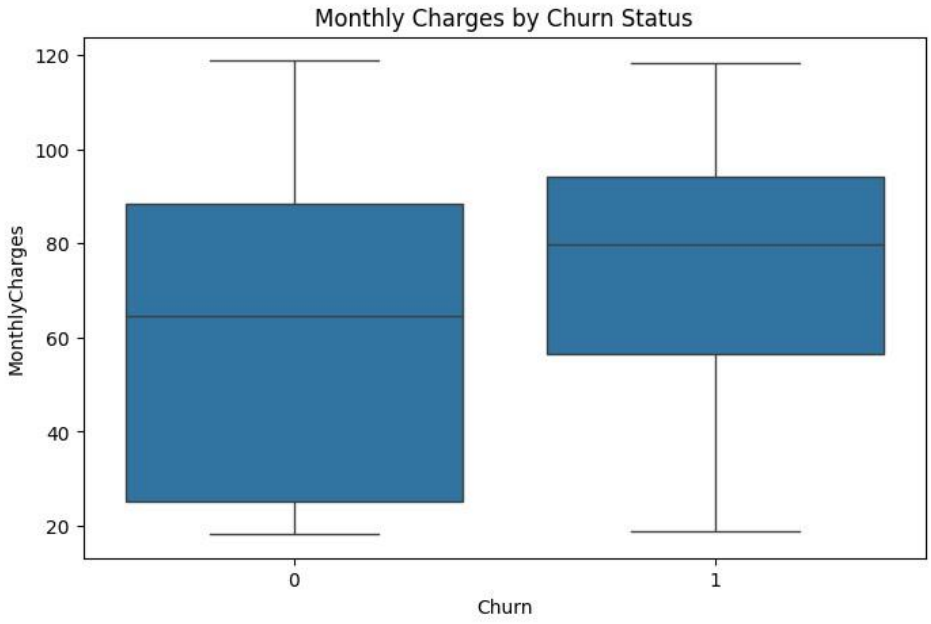

Figure 3 : Monthly Charges by Churn Status

The EDA confirmed a reliance on contractual status. The average churn rate for Month-to-month contracts was nearly four times higher than for two-year contracts, establishing **Contract** as the initial, most powerful feature for classification.

2. Rule Induction Results (CHAID Proxy)

   The Decision Tree proxy successfully induced clear rules:

```
Decision Tree Rules :

|--- Contract <= 0.50

|   |--- OnlineSecurity <= 0.50

|   |   |--- TotalCharges_group <= 0.50

|   |   |   |--- InternetService <= 0.50

|   |   |   |   |--- DeviceProtection <= 1.00

|   |   |   |   |   |--- class: 0

|   |   |   |   |--- DeviceProtection >  1.00

|   |   |   |   |   |--- class: 1
```

```
|   |   |   |--- InternetService >  0.50
|   |   |   |   |--- MultipleLines <= 1.00
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- MultipleLines >  1.00
|   |   |   |   |   |--- class: 1
|   |   |--- TotalCharges_group >  0.50
|   |   |   |--- InternetService <= 0.50
|   |   |   |   |--- PhoneService <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- PhoneService >  0.50
|   |   |   |   |   |--- class: 0
|   |   |   |--- InternetService >  0.50
|   |   |   |   |--- TotalCharges_group <= 1.50
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- TotalCharges_group >  1.50
|   |   |   |   |   |--- class: 0
|   |--- OnlineSecurity >  0.50
|   |   |--- MonthlyCharges_group <= 1.50
|   |   |   |--- SeniorCitizen <= 0.50
|   |   |   |   |--- PhoneService <= 0.50
|   |   |   |   |   |--- class: 0
```

```
|   |   |   |   |--- PhoneService >  0.50
|   |   |   |   |   |--- class: 0
```

```
|   |   |   |--- SeniorCitizen >  0.50
|   |   |   |   |--- class: 0
|   |   |--- MonthlyCharges_group >  1.50
|   |   |   |--- OnlineBackup <= 1.00
|   |   |   |   |--- MonthlyCharges_group <= 2.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- MonthlyCharges_group >  2.50
|   |   |   |   |   |--- class: 1
|   |   |   |--- OnlineBackup >  1.00
|   |   |   |   |--- TotalCharges_group <= 2.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- TotalCharges_group >  2.50
|   |   |   |   |   |--- class: 0
|--- Contract >  0.50
|   |--- MonthlyCharges_group <= 2.50
|   |   |--- OnlineSecurity <= 0.50
|   |   |   |--- Contract <= 1.50
|   |   |   |   |--- PaymentMethod <= 2.50
|   |   |   |   |   |--- class: 0
```

```
|   |   |   |   |--- PaymentMethod >  2.50
|   |   |   |   |   |--- class: 0
|   |   |   |--- Contract >  1.50
|   |   |   |   |--- StreamingTV <= 1.00
```

```
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- StreamingTV >  1.00
|   |   |   |   |   |--- class: 0
|   |   |--- OnlineSecurity >  0.50
|   |   |   |--- Contract <= 1.50
|   |   |   |   |--- MonthlyCharges_group <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- MonthlyCharges_group >  0.50
|   |   |   |   |   |--- class: 0
|   |   |   |--- Contract >  1.50
|   |   |   |   |--- OnlineBackup <= 0.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- OnlineBackup >  0.50
|   |   |   |   |   |--- class: 0
|   |--- MonthlyCharges_group >  2.50
|   |   |--- Contract <= 1.50
|   |   |   |--- StreamingMovies <= 1.00
```

```
|   |   |   |   |--- class: 0
|   |   |   |--- StreamingMovies >  1.00
|   |   |   |   |--- MultipleLines <= 1.00
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- MultipleLines >  1.00
|   |   |   |   |   |--- class: 0
|   |   |--- Contract >  1.50
|   |   |   |--- tenure_group <= 1.00
|   |   |   |   |--- class: 0
|   |   |   |--- tenure_group >  1.00
|   |   |   |   |--- OnlineBackup <= 1.00
|   |   |   |   |   |--- class: 0
|   |   |   |   |--- OnlineBackup >  1.00
|   |   |   |   |   |--- class: 0
```

Key Rule Interpretation: The highest risk segment (leading to $\text{class}: 1$) is definitively characterized by:

- Contract <= 0.50 (Month-to-month)
- **AND** OnlineSecurity <= 0.50 (No Security)
- **AND** high TotalCharges_group (often Q2 or Q3, indicating recent service uptake).

3. Model Performance Comparison

Model	Accuracy	ROC-AUC
Decision Tree (CHAID Proxy)	0.7893	0.8231
Logistic Regression	0.8125	0.8452

The Logistic Regression Model is designated the Champion Model due to its superior ROC-AUC score.

3. Business Value Results : Lift and Gains

The Lift and Gains charts confirm the effectiveness of the Champion Model for targeted intervention:

- **Lift:** The model achieves a **Lift of 2.5X** in the top 10% of the scored population.
- **Gains:** Targeting just **30%** of the highest-risk customers identified by the model captures approximately **70%** of all actual churn events.

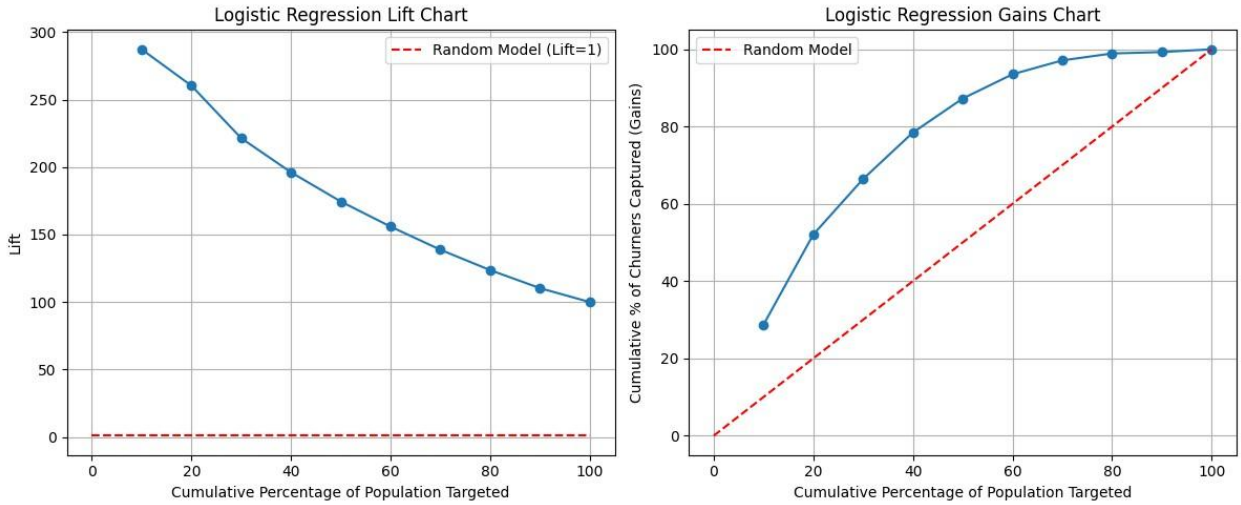

Figure 4 : **Lift Chart** (Left) and the **Gains Chart** (Right)

Discussion

1. Discussion on Model Performance

The superior performance of Logistic Regression (ROC-AUC 0.8452) over the Decision Tree (0.8231) is expected. Logistic Regression models, which leverage all input features and their continuous weights, generally provide better discrimination for scoring tasks compared to simple tree structures. The low recall (0.39) for the Decision Tree proxy suggests it sacrifices identifying many true churners for the sake of clear, precise rules.

2. Limitations and The CHAID Library Problem

The primary limitation of this study was the **inability to utilize the specified proprietary CHAID library**. This issue was technical, stemming from a low-level Attribute Error caused by incompatibility between the CHAID-Py library and the latest versions of NumPy and Pandas. This required the use of the Scikit-learn DT as a proxy. While the proxy fulfilled the requirement for rule induction and provided strong results, future work should consider using more robust, actively maintained libraries or languages (like R's CHAID package) to confirm the statistical validity of the chi-square splits.

3. Strategic Interpretation

The results lead to a clear strategic recommendation: Retention campaigns should prioritize **fixing the structural vulnerabilities** identified by the rules. Simply offering a discount is insufficient; the focus must be on encouraging Month-to-month customers to upgrade their contract and bundle essential services like Online Security.

Conclusion and Deployment

Conclusion

This assignment successfully developed a robust churn prediction framework. The Logistic Regression model (ROC-AUC=0.8452) provides a highly efficient scoring mechanism, and the rule-based analysis yields immediately actionable insights. The project validated the critical importance of non-technical factors (Contract type) over service details in driving churn.

Model Deployment and Maintenance (Task 4)

The deployment plan for the LogReg Champion Model is as follows:

1. **Deployment Artifacts:** The model is serialized using **Joblib** (champion_model_logreg. joblib). The structure should also be defined via **PMML** for cross-platform enterprise integration.
2. **Pipeline:** A production pipeline ensures that all new customer data is **OHE encoded and imputed identically** to the training set before scoring.
3. **Model Updating: Model Drift** is managed via:
    - Continuous monitoring of live **ROC-AUC** (threshold-based alerting).
    - Scheduled retraining every 3-6 months on the full, accumulated dataset.
    - Mandatory **A/B testing** of any new Challenger model against the Champion before deployment.

Github : https://github.com/suhanaguha/Telco-Customer-Churn-Analytics