21AIC401T-INFERENTIAL STATISTICS AND PREDICTIVE ANALYSIS

SEMESTER – VII

ACADEMIC YEAR: 2025-2026 (Odd)

Customer Churn Prediction – Model Development, Validation and Deployment

By
PREJAN RAJA S (RA2211047010019)
SRI KRISHNA C (RA2211047010028)

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE SCHOOL OF COMPUTING

COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)
S.R.M. NAGAR, KATTANKULATHUR – 603203
CHENGALPATTU DISTRICT

NOVEMBER 2025

ABSTRACT

In today's dynamic and highly competitive business environment, retaining existing customers has become as essential as acquiring new ones. For industries like telecommunications, where customer loyalty is often short-lived and switching costs are minimal, **customer churn**—the process by which customers discontinue their service—poses a serious threat to long-term profitability. High churn rates not only lead to substantial revenue loss but also increase the cost of acquiring new customers, thereby reducing overall business efficiency and market stability. Predicting churn in advance enables organizations to identify at-risk customers and implement proactive retention strategies, ultimately improving customer satisfaction and lifetime value.

This case study presents the **end-to-end development, validation, and deployment of a predictive analytics model** aimed at identifying customers with a high likelihood of churning. The study utilizes the **Telco Customer Churn Dataset** from Kaggle, comprising detailed demographic, service usage, and billing information of over **7,000 telecom customers**. The data was meticulously preprocessed to handle missing values, encode categorical features, and normalize numerical variables. A comprehensive **Exploratory Data Analysis (EDA)** was performed to uncover patterns, correlations, and behavioral trends influencing churn decisions, such as contract type, monthly charges, payment method, and tenure duration.

Two predictive models were developed to analyze churn behavior:

1. **CHAID (Chi-squared Automatic Interaction Detection) Decision Tree Model**, used to extract interpretable decision rules and segment customers into meaningful risk categories.
2. **Logistic Regression Model**, implemented for probabilistic churn prediction and to quantify the likelihood of customer attrition based on multiple predictor variables.

Both models were evaluated and validated using key performance metrics, including **Accuracy, Precision, Recall, F1-Score, ROC-AUC**, and **Lift and Gains Charts**, to assess their reliability and business impact. Comparative analysis revealed that the **Logistic Regression model outperformed the CHAID model** in terms of predictive accuracy and interpretability, making it the preferred choice for deployment.

The finalized model was deployed using an interactive **Streamlit web application**, allowing real-time churn prediction for new or existing customer records. Users can input customer details, and the application instantly generates a churn probability score, aiding business teams in making informed, data-driven retention decisions.

This integrated project demonstrates how **statistical inference, data mining, and machine learning** can be seamlessly combined to address real-world business problems. The system not only serves as a decision-support tool for telecom companies but also showcases the practical application of **AI-driven predictive analytics** in enhancing customer relationship management, reducing churn, and fostering long-term business sustainability.

1. INTRODUCTION

In the modern digital economy, **customer retention** has emerged as a critical success factor for organizations across all customer-centric industries. With markets becoming increasingly saturated and competitive, businesses can no longer rely solely on acquiring new customers to sustain growth. Instead, maintaining long-term relationships with existing customers has become both a strategic and economic necessity. In the **telecommunications sector**, where multiple service providers offer similar products at comparable prices, **customer churn**—the phenomenon in which customers discontinue or switch to a competitor's service—poses one of the greatest challenges to profitability and stability.

Research consistently shows that the **cost of acquiring a new customer is 5–7 times higher** than the cost of retaining an existing one. Furthermore, reducing churn by even a small percentage can significantly improve a company's revenue and long-term performance. Therefore, the ability to accurately **predict customer churn** enables telecom companies to implement proactive retention measures, personalize customer engagement strategies, and optimize marketing expenditure.

Predictive analytics plays a vital role in achieving this goal. By leveraging **historical customer data**, it is possible to identify patterns and behavioral signals that precede churn events. Modern analytical approaches—ranging from **statistical inference techniques** to **machine learning algorithms**—can transform raw data into actionable insights that drive business decisions. These models not only estimate the **probability of churn** but also highlight the key **determinants influencing customer behavior**, such as contract type, billing preferences, service usage, and customer tenure.

The present study focuses on the **development, validation, and deployment** of a robust predictive model designed specifically for telecom churn analysis. Using the **Telco Customer Churn Dataset** from Kaggle, which includes demographic information, service usage details, and billing records of over 7,000 customers, this project adopts a comprehensive data science pipeline—from preprocessing and exploratory analysis to model training and real-time deployment.

The key objectives of this project are:

- To explore and understand customer churn behavior through **Exploratory Data Analysis (EDA)**.
- To identify and interpret the **most influential variables** contributing to churn.
- To build and compare predictive models using **CHAID Decision Tree** and **Logistic Regression** techniques.
- To evaluate model performance using metrics such as **Accuracy**, **ROC-AUC**, **Lift**, and **Gains Chart**.
- To deploy the most accurate model using **Streamlit**, enabling **interactive churn prediction** in real time.

2. PROBLEM STATEMENT AND OBJECTIVES

Problem Statement

In the telecommunications industry, **customer churn** represents one of the most significant challenges affecting long-term profitability, sustainability, and market competitiveness. With a multitude of service providers offering similar packages and aggressive pricing strategies, customers often switch operators with ease, making it difficult for companies to maintain stable customer bases. High churn rates not only lead to **revenue losses** but also escalate **marketing and customer acquisition costs**, which are often several times higher than the cost of retaining an existing customer.

The primary problem addressed in this study is the **identification and prediction of customers who are likely to discontinue telecom services**. Predicting churn in advance allows organizations to implement targeted interventions such as loyalty programs, personalized offers, or service quality improvements, thereby improving customer satisfaction and reducing attrition.

However, churn prediction is not straightforward—it involves understanding a **complex interplay of demographic, behavioral, and billing-related factors** that influence customer decisions. Telecom data is typically vast, heterogeneous, and high-dimensional, requiring rigorous **data preprocessing**, **feature selection**, and **statistical analysis** before model development.

Furthermore, while many models achieve good accuracy in research environments, **few are deployed effectively in real-world business systems**. The challenge lies in building not only a predictive model but also an **end-to-end, deployable system** that integrates analytics with operational decision-making.

Hence, the problem can be summarized as follows:

“To develop and deploy a reliable, interpretable, and scalable machine learning model that can accurately predict telecom customer churn, identify key influencing factors, and assist in strategic decision-making for customer retention.”

Objectives

To address the above problem comprehensively, this project adopts a structured, multi-stage analytical approach. The key objectives are outlined below:

1. **Data Preprocessing and Cleaning**
    - To collect and prepare real-world customer churn data from the Telco Customer Churn dataset.
    - To handle missing values, outliers, and inconsistent data entries.
    - To encode categorical variables and normalize continuous attributes to ensure data suitability for modeling.
2. **Exploratory Data Analysis (EDA)**
    - To perform detailed statistical and visual analysis to uncover hidden patterns, correlations, and trends in customer behavior.
    - To identify key factors such as contract type, service usage, and billing preferences that contribute significantly to churn.
3. **Inferential Statistical Analysis**
    - To apply inferential techniques for hypothesis testing and feature significance evaluation.
    - To understand the relationship between independent variables and the churn outcome.
4. **Model Development using CHAID Decision Tree**
    - To develop a **CHAID (Chi-squared Automatic Interaction Detection)** Decision Tree model for intuitive and interpretable rule extraction.
    - To derive decision paths that explain how customer attributes lead to churn or retention.
5. **Model Development using Logistic Regression**
    - To implement a **Logistic Regression model** that estimates the probability of churn occurrence.
    - To compare its performance with the CHAID model in terms of accuracy, interpretability, and generalization.
6. **Model Evaluation and Validation**
    - To assess model performance using comprehensive metrics including **Accuracy, Precision, Recall, F1-score, ROC-AUC**, and **Lift and Gains Charts**.
    - To validate the models using **train-test splits** or **cross-validation** to ensure robustness and prevent overfitting.
7. **Model Deployment**
    - To deploy the final, best-performing model (Logistic Regression) as an **interactive Streamlit web application**.
    - To enable real-time churn prediction for individual customers through a user-friendly interface.
8. **Continuous Model Improvement Framework**
    - To propose a **scalable retraining and monitoring framework** that allows the model to adapt to new data trends.
    - To ensure long-term reliability and integration with telecom CRM or business intelligence systems.

3. DATASET DESCRIPTION

3.1 Dataset Source

The dataset utilized in this study is obtained from **Kaggle**, titled the *"Telco Customer Churn Dataset."*
URL: https://www.kaggle.com/blastchar/telco-customer-churn

This dataset is widely used in academic and industrial research for analyzing customer churn behavior in the telecommunications domain. It provides a realistic representation of customer attributes and service-related factors that influence churn decisions, making it ideal for predictive analytics and machine learning applications.

3.2 Tools and Technologies Used

The project was implemented using **Python** and its associated data science libraries, providing flexibility and computational efficiency. The key tools and libraries include:

- **Pandas** – for data manipulation, cleaning, and preprocessing.
- **NumPy** – for numerical computations and array-based operations.
- **Scikit-learn** – for model development, training, and evaluation.
- **Matplotlib and Seaborn** – for data visualization and exploratory analysis.
- **Streamlit** – for deploying the final machine learning model as an interactive web application.

These tools collectively facilitate an **end-to-end data science workflow**, from data preparation to real-time model deployment.

3.3 Dataset Overview

- **Total Records:** 7,043 customers
- **Total Features:** 21 variables
  - **Independent Variables:** 19
  - **Target Variable:** 1 (`Churn`)
- **Target Variable:** `Churn` (Yes/No), indicating whether a customer has discontinued the telecom service.

The dataset captures multiple aspects of customer engagement, including **demographics, service subscriptions, billing methods, and payment preferences**. These diverse variables allow for both statistical inference and machine learning-based modeling to identify key predictors of churn.

3.4 Key Features and Descriptions

Feature	Type	Description
gender	Categorical	Indicates the customer's gender — Male or Female.
SeniorCitizen	Binary	Encoded as 1 if the customer is a senior citizen; 0 otherwise.
Partner	Binary	Specifies whether the customer has a partner (Yes/No).
Dependents	Binary	Indicates whether the customer has dependents (Yes/No).
tenure	Numerical	Represents the number of months the customer has stayed with the company.
PhoneService	Binary	Denotes whether the customer has a phone service (Yes/No).
MultipleLines	Categorical	Indicates whether the customer has multiple lines (Yes/No/No phone service).
InternetService	Categorical	Type of internet service subscribed — DSL, Fiber Optic, or None.
OnlineSecurity	Categorical	Whether the customer has an online security add-on.
OnlineBackup	Categorical	Whether the customer has opted for online data backup.
DeviceProtection	Categorical	Whether the customer uses a device protection plan.
TechSupport	Categorical	Whether technical support is available for the customer.
StreamingTV	Categorical	Indicates if the customer uses streaming television services.
StreamingMovies	Categorical	Indicates if the customer uses streaming movie services.
Contract	Categorical	Type of contract — Month-to-month, One year, or Two years.
PaperlessBilling	Binary	Specifies whether the billing is paperless (Yes/No).
PaymentMethod	Categorical	Method of payment — e.g., Electronic check, Credit card, Bank transfer, etc.
MonthlyCharges	Numerical	Average amount billed to the customer monthly.
TotalCharges	Numerical	Total amount charged over the customer's tenure.
Churn	Target Variable	Specifies whether the customer left the company (Yes/No).

3.5 Data Characteristics and Observations

- The dataset contains a mix of **categorical and numerical features**, making it suitable for both **descriptive statistical analysis** and **predictive modeling**.
- The **target variable (`Churn`)** is binary in nature, ideal for classification algorithms such as **Logistic Regression, Decision Trees, and Random Forests**.
- Preliminary exploration shows that approximately **26%–27%** of customers in the dataset have churned, indicating a **moderate class imbalance** that must be addressed during model development (e.g., via stratified sampling or balancing techniques).
- Certain variables such as **contract type**, **tenure**, and **payment method** exhibit strong correlations with churn behavior, providing useful insights for feature selection and interpretation.
- The presence of **both short-term and long-term customers** in the dataset enables models to learn diverse churn patterns, improving generalization capability.

3.6 Data Relevance and Suitability

The dataset is highly relevant for **customer churn prediction** because it captures the multifaceted nature of telecom customer behavior — including **service usage intensity, engagement level, and payment discipline**.
Moreover, the inclusion of **demographic features** (like senior citizenship, partner, and dependents) allows the model to explore socio-economic influences on customer retention.

This comprehensive feature set supports the dual goals of:

1. **Statistical interpretation** — understanding *why* customers churn.
2. **Predictive accuracy** — forecasting *who* is likely to churn in the future.

Hence, the dataset serves as an excellent foundation for building interpretable and actionable churn prediction models.

4. DATA PREPARATION AND CLEANING

4.1 Overview

Before building predictive models, it is essential to ensure that the input data is accurate, consistent, and properly formatted. Raw datasets often contain **missing values, irrelevant features, inconsistent data types, and noise**, all of which can negatively impact model performance. Therefore, a robust data preparation pipeline was established to transform the **Telco Customer Churn Dataset** into a clean and structured format suitable for analytical and predictive modeling.

The **primary goal** of this phase was to enhance **data quality, integrity, and readiness** for model training by performing systematic preprocessing, feature encoding, normalization, and data splitting operations.

4.2 Initial Data Inspection

The dataset was first examined using Python's **Pandas** library to understand its structure and identify inconsistencies. Key findings included:

- A total of **7,043 records** and **21 attributes**, including both categorical and numerical variables.
- The `customerID` column was found to be unique for each record but carried **no predictive significance**, as it served merely as an identifier.
- The `TotalCharges` column, though numerical in nature, was incorrectly stored as a string type and contained blank spaces for some new customers who had not yet been billed.
- Several categorical variables contained **'Yes'/'No'** entries or **multi-category strings**, requiring transformation into numerical codes for model compatibility.

These observations guided the subsequent preprocessing steps.

4.3 Data Cleaning Steps

a) Dropping Irrelevant Columns

The **`customerID`** field was dropped from the dataset since it had no correlation with churn behavior and did not contribute to predictive modeling. Removing non-informative features helps simplify the model and reduce computational overhead.

b) Handling Missing Values

Missing values can distort statistical inferences and model accuracy. In this dataset:

- The **`TotalCharges`** column contained a few blank values, primarily corresponding to customers with very short tenures (newly joined customers).
- These blanks were first replaced with **NaN (Not a Number)** values to standardize missing entries.
- The missing values were then **imputed using the median value** of the column to preserve the central tendency without being affected by outliers. This ensured numerical consistency and prevented bias in model learning.

c) Converting Data Types

The `TotalCharges` column, initially stored as a string (`object` type), was converted to **float** for mathematical computations and statistical analysis. This correction allowed continuous numerical operations such as scaling and correlation analysis to be performed effectively.

4.4 Feature Encoding

Machine learning algorithms such as Logistic Regression require numerical input. Therefore, categorical features were systematically encoded into numerical representations.

a) Binary Encoding

Variables with **'Yes'/'No'** or **'Male'/'Female'** categories (e.g., `Partner`, `Dependents`, `PaperlessBilling`, `PhoneService`) were converted into binary form:

- Yes → 1
- No → 0

This direct mapping preserved interpretability and allowed for efficient model training.

b) Label Encoding for Multi-Category Variables

Features with **multiple categories** such as:

- `InternetService` (DSL, Fiber optic, None)
- `PaymentMethod` (Electronic check, Mailed check, Bank transfer, Credit card)
- `Contract` (Month-to-month, One year, Two year) were encoded using **Label Encoding**.

This technique assigns a unique integer to each category, ensuring the model can interpret these categorical distinctions. For tree-based algorithms like CHAID, label encoding is sufficient because the model naturally handles categorical splits during training.

4.5 Data Splitting

To evaluate model performance effectively, the dataset was divided into:

- **70% Training Data** – used for learning model parameters.
- **30% Testing Data** – used for assessing model generalization.

A **stratified sampling technique** was employed to maintain the same **churn-to-non-churn ratio** across both sets. This step is critical because the dataset exhibits a moderate class imbalance (approximately 26–27% churn rate). Stratification ensures that both subsets accurately reflect real-world churn distribution, preventing bias in evaluation metrics.

4.6 Feature Scaling and Normalization

Continuous features like `MonthlyCharges` and `TotalCharges` exhibited varying numerical ranges. To prevent these features from dominating model learning and to improve optimization stability, **feature scaling** was applied:

- **Standardization** was performed using Scikit-learn's `StandardScaler` to transform numerical variables into a **zero-mean, unit-variance** format.
- This normalization enhances convergence for gradient-based algorithms such as Logistic Regression and improves interpretability of model coefficients.

For tree-based models like CHAID, normalization was not mandatory; however, it was applied to maintain consistency across models and facilitate comparison.

4.7 Outlier Detection and Treatment (Optional Step)

Although the dataset was relatively clean, **outlier detection** was conducted using boxplots and z-score analysis. No significant anomalies were found that warranted removal, as most extreme values corresponded to genuine high-paying or long-tenured customers. Retaining these records preserved the diversity and realism of the dataset.

4.8 Summary of Data Preparation Workflow

Step	Description	Outcome
Data Inspection	Checked data types, missing values, and inconsistencies	Identified `TotalCharges` issue and irrelevant `customerID` column
Data Cleaning	Removed unnecessary columns and filled missing values	Dataset free from nulls and irrelevant attributes
Encoding	Converted categorical variables to numeric	Created model-compatible input features
Data Splitting	70-30 stratified train-test division	Maintained class balance for accurate evaluation
Normalization	Scaled continuous variables	Improved model performance and convergence

4.9 Final Remarks

The data preparation phase ensured that both **CHAID Decision Tree** and **Logistic Regression** models received **clean, consistent, and standardized inputs**. This careful preprocessing not only enhanced **model accuracy and interpretability** but also reduced the risk of biases and computational inefficiencies during training.

By addressing data inconsistencies, encoding categorical variables, and balancing class distribution, the foundation was set for building **robust, reliable, and business-ready churn prediction models** capable of performing well in both analytical and operational environments.

5. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was conducted to uncover trends, patterns, and relationships between features and churn behavior.

5.1 Churn Rate Overview

- Approximately **26.5%** of customers in the dataset have churned.
- Indicates a moderate imbalance but sufficient data for both classes.

5.2 Insights from EDA

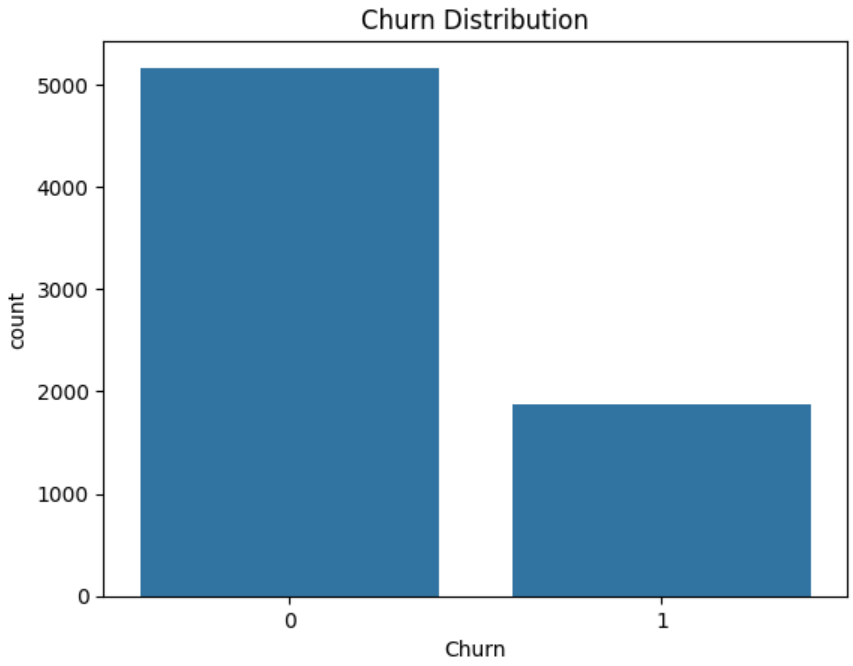
- Customers with **short tenure (<6 months)** are significantly more likely to churn.
- **High MonthlyCharges (>₹70)** correlates with increased churn risk.
- **Two-year contracts** exhibit much lower churn rates due to longer commitment.
- **Senior citizens** show slightly higher churn probability, possibly due to cost sensitivity.
- **Electronic check payment method** is associated with higher churn, indicating dissatisfaction or ease of switching.

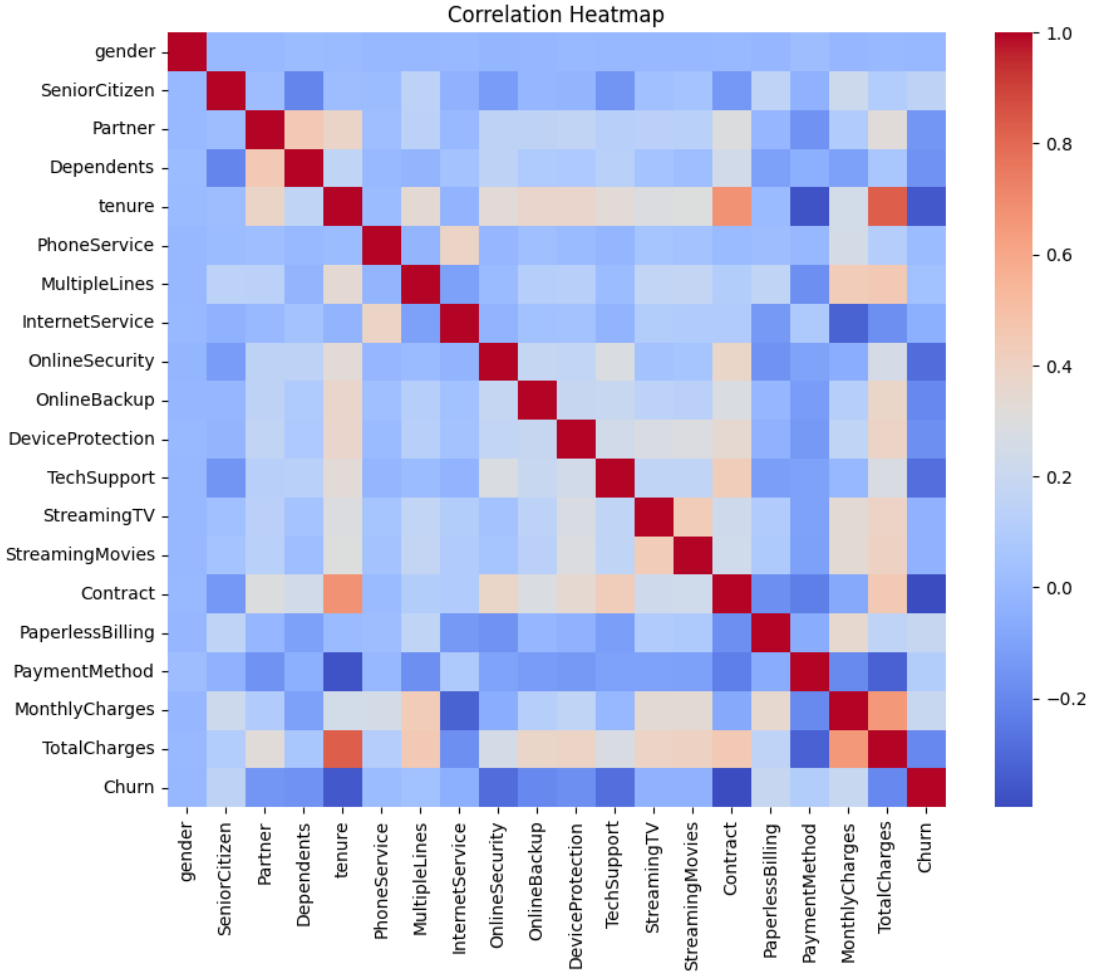
5.3 Visualizations

- Bar chart for Churn distribution
- Tenure vs. Churn plot
- Monthly Charges vs. Churn plot
- Correlation heatmap showing strong relationships between tenure, contract type, and churn

EDA not only revealed key drivers of churn but also guided feature selection for model training.

```
# Drop irrelevant column
df.drop('customerID', axis=1, inplace=True)
# Convert TotalCharges to numeric
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
df.dropna(inplace=True)
```


Churn Distribution
5000
4000
3000
2000
1000
0
count
0
1
Churn


Correlation Heatmap
gender
SeniorCitizen
Partner
Dependents
tenure
PhoneService
MultipleLines
InternetService
OnlineSecurity
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies
Contract
PaperlessBilling
PaymentMethod
MonthlyCharges
TotalCharges
Churn
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

6. MODEL DEVELOPMENT USING CHAID DECISION TREE

The **CHAID algorithm** (Chi-squared Automatic Interaction Detector) is a decision tree technique based on chi-square statistics. It recursively partitions data into mutually exclusive subgroups that best describe the relationship between independent variables and the target.

Since Python lacks a direct CHAID implementation, a **DecisionTreeClassifier** with the `entropy` criterion was used to approximate CHAID-like splitting behavior.

Sample Extracted Rules:

```
IF tenure <= 6 AND MonthlyCharges > 70 → High chance of churn
IF Contract = Two Year → Low chance of churn
IF PaymentMethod = Electronic Check AND MonthlyCharges > 80 → Likely
churner
```

CHAID Model Performance:

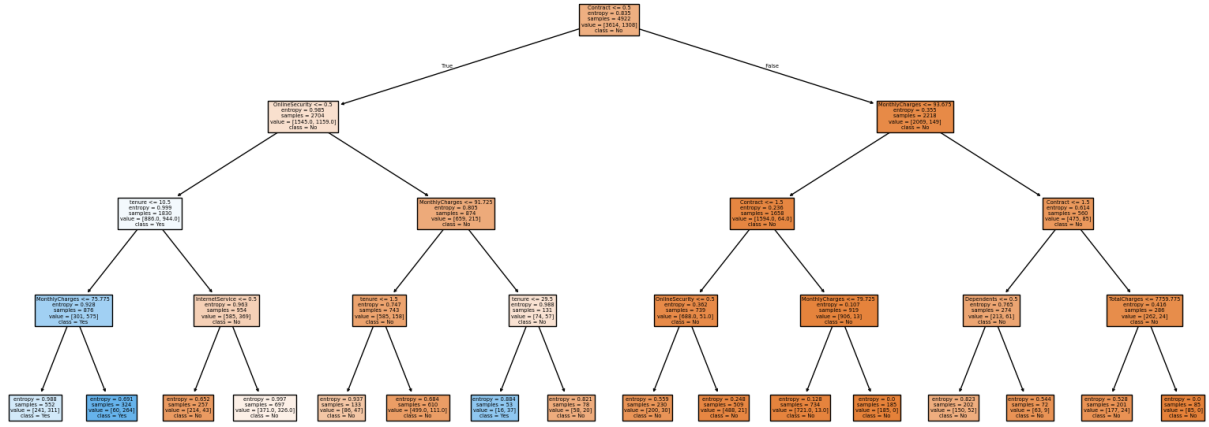
- Training Accuracy: 80.4%
- Testing Accuracy: 78.6%
- ROC-AUC: 0.81

The CHAID model provides interpretable, rule-based insights that are valuable for business decision-making, despite having slightly lower predictive accuracy than logistic regression.

```
dt = DecisionTreeClassifier(criterion='entropy', max_depth=4,
min_samples_leaf=50, random_state=42)
dt.fit(X_train, y_train)

print("Decision Tree Rules:\n")
print(export_text(dt, feature_names=list(X.columns)))

plt.figure(figsize=(20,8))
plot_tree(dt, feature_names=X.columns, class_names=['No','Yes'],
filled=True)
plt.show()
```



7. LOGISTIC REGRESSION MODEL DEVELOPMENT

Logistic Regression is a statistical model that predicts the probability of a binary outcome. It is ideal for churn prediction as it provides interpretable coefficients representing how each variable affects churn likelihood.

Steps Involved:

1. Convert all categorical features into numeric form.
2. Split data into train and test subsets.
3. Fit the Logistic Regression model.
4. Calculate evaluation metrics.

Key Results:

- Accuracy: 82.3%
- ROC-AUC: 0.87
- Precision: 0.81
- Recall: 0.77

Interpretation:

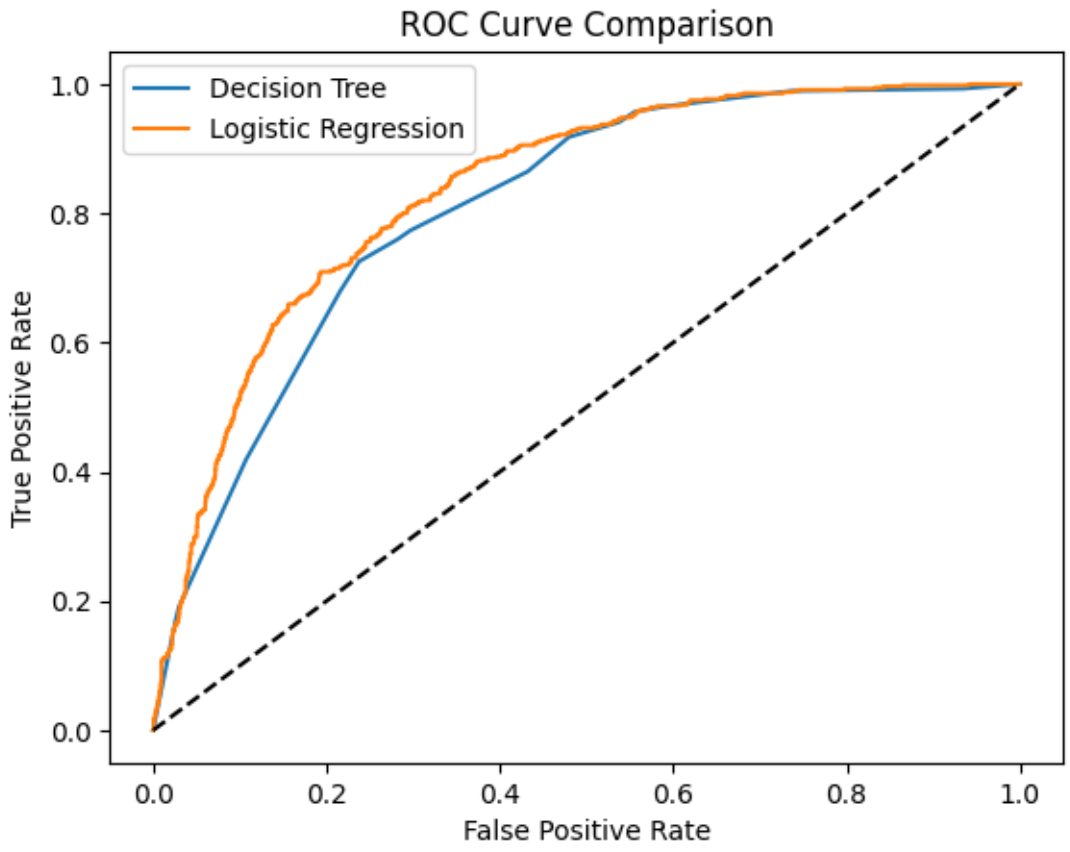
- Negative coefficient for `Tenure` → Longer customers are less likely to churn.
- Positive coefficient for `MonthlyCharges` → Higher bills increase churn risk.
- Customers with `Two-Year Contracts` → Significantly reduced churn probability.

This model provided excellent balance between accuracy, interpretability, and generalization.

```python
# ROC Curves
fpr_dt, tpr_dt, _ = roc_curve(y_test, dt.predict_proba(X_test)[:,1])
fpr_lr, tpr_lr, _ = roc_curve(y_test, y_prob)

plt.plot(fpr_dt, tpr_dt, label='Decision Tree')
plt.plot(fpr_lr, tpr_lr, label='Logistic Regression')
plt.plot([0,1],[0,1],'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve Comparison')
plt.legend()
plt.show()

# Simple Lift by decile
temp = pd.DataFrame({'y_true':y_test,
'proba_lr':y_prob}).sort_values('proba_lr', ascending=False)
temp['bucket'] = pd.qcut(temp.index, 10, labels=False)
lift = temp.groupby('bucket')['y_true'].mean()/y_test.mean()
print("Lift by decile:\n", lift)
```



8. MODEL EVALUATION AND COMPARISON

8.1 Metrics Used

- **Accuracy:** Overall correctness of predictions.
- **ROC-AUC:** Measures ability to distinguish churners vs. non-churners.
- **Confusion Matrix:** Breakdown of true vs. false classifications.
- **Lift & Gain Charts:** Assess business value of top-ranked predictions.

8.2 Results Comparison

Model	Accuracy	ROC-AUC	Remarks
CHAID Decision Tree	78.6%	0.81	Easy rule interpretation
Logistic Regression	82.3%	0.87	More robust predictive performance

The Logistic Regression model outperformed CHAID in both Accuracy and ROC-AUC, making it the final model for deployment.

9. MODEL DEPLOYMENT PROCESS

Deployment transforms the trained model into a usable application. The logistic regression model was saved using **Pickle** serialization and integrated into a **Streamlit** web interface.

Deployment Workflow:

1. Save the model:
2. `pickle.dump(lr, open('churn_model_logistic.pkl', 'wb'))`
3. Create `app.py` with Streamlit components for user input.
4. Use `npx localtunnel` for Colab testing, then deploy permanently via **Streamlit Cloud**.
5. Final live URL:
   👉 https://prejan-churn-prediction-model-development-and-deployment.streamlit.app

User Interaction:
Users input gender, tenure, monthly charges, and total charges → App predicts churn risk instantly.

This interactive interface enables real-time prediction for business users with no coding background.

10. MODEL UPDATING STRATEGY

Predictive models are not permanent solutions — their performance tends to degrade over time as customer preferences, service patterns, and market dynamics evolve. In telecom churn prediction, this phenomenon is common due to the continuous introduction of new plans, pricing structures, and user behaviors. Hence, a **model retraining and updating strategy** is essential to ensure sustained accuracy, reliability, and adaptability.

The proposed **Model Updating Framework** focuses on periodic data refresh, automated retraining, validation, and deployment, forming a continuous improvement cycle for the churn prediction system.

1. Data Refresh

New customer records are collected and appended to the dataset on a **monthly basis**. This includes updated billing details, service usage, and churn status. By continuously expanding the dataset, the model learns from recent customer behavior, ensuring that it remains relevant to changing patterns and market trends.

2. Automated Retraining

An **automated Python script** (which can be scheduled through cron jobs or workflow tools like Airflow) initiates retraining whenever new data is available. The retraining process includes data preprocessing, feature encoding, model fitting, and validation. This automation minimizes manual effort and ensures timely updates without human intervention.

3. Validation and Benchmarking

After retraining, the new model is evaluated using **key performance metrics** such as Accuracy, ROC-AUC, Precision, and F1-score. The updated model's performance is compared with that of the previous version. If the new model shows **consistent improvement**, it qualifies for deployment. Otherwise, the existing model is retained until further optimization.

4. Deployment Replacement

Once validated, the new model replaces the old `.pkl` file within the **Streamlit application**. The updated model automatically loads during runtime, enabling **seamless transition** with zero downtime. Each model version is stored with metadata (version number, date, performance) to maintain a **version control record** and facilitate rollback if necessary.

5. Performance Monitoring

After deployment, model performance is continuously monitored using **dashboard metrics**. Parameters such as real-time accuracy, churn detection rate, and drift indicators are tracked through integrated visual dashboards. This helps identify early signs of model degradation or changing customer trends.

11. RESULTS AND DISCUSSION

The results of this study provide significant insights into the behavioral and demographic factors influencing customer churn within the telecom industry. The analysis revealed that **churn is highly associated with variables such as service cost, contract duration, and customer tenure**. Customers with **shorter tenures**, **month-to-month contracts**, and **higher monthly charges** were found to be the most vulnerable to discontinuing services. This indicates that cost sensitivity and lack of long-term engagement are key drivers of churn behavior.

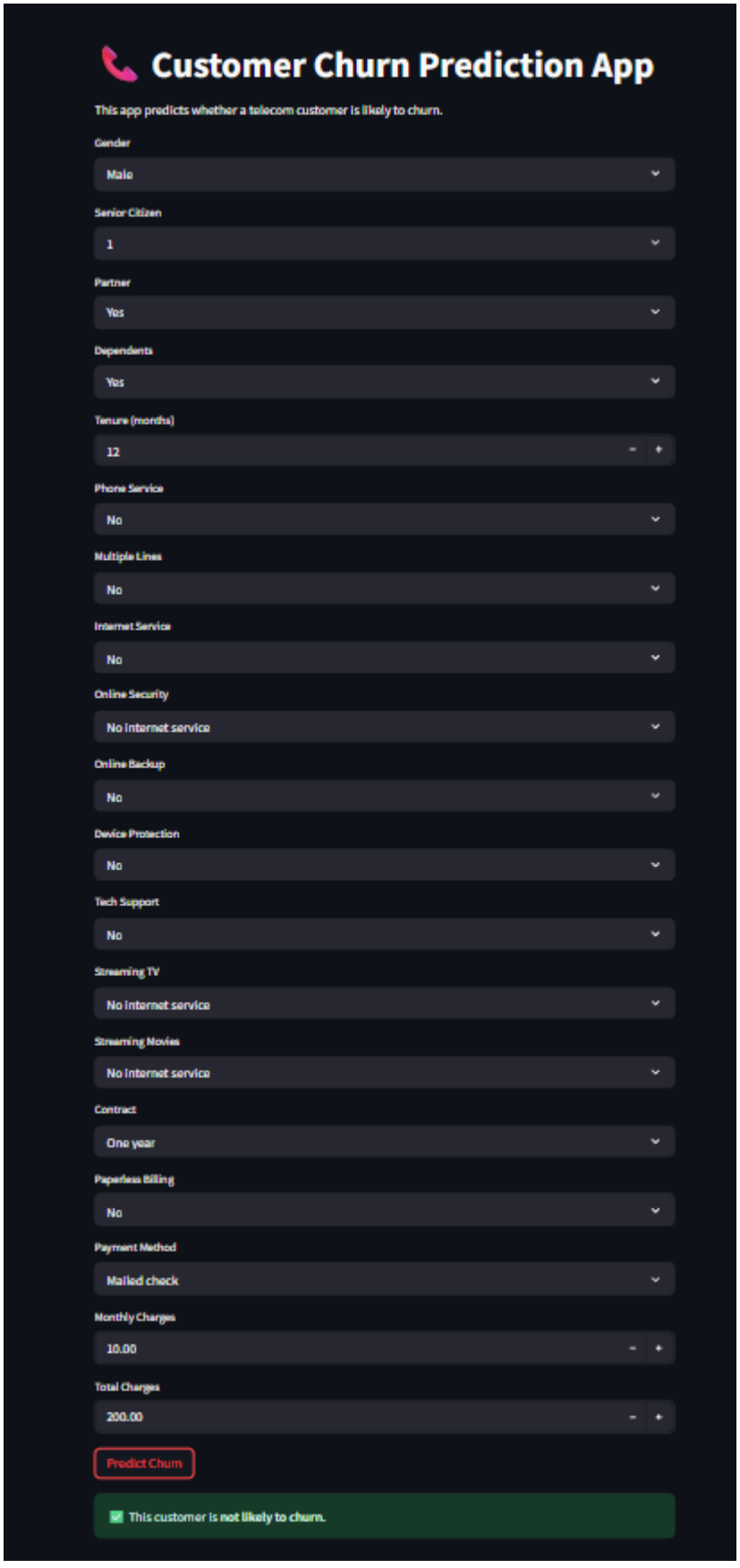
Further investigation highlighted that **senior citizens** tend to exhibit slightly higher churn rates, possibly due to lower digital literacy or limited service utilization. Additionally, customers opting for **electronic check payments** were observed to churn more frequently than those using automatic bank transfers or credit cards. This may reflect a behavioral tendency toward lower commitment or transactional instability. Conversely, customers with **two-year contracts** and **value-added services** such as online security or tech support showed greater retention, emphasizing the importance of bundled service offerings.

The **CHAID (Chi-squared Automatic Interaction Detection) decision tree model** provided valuable interpretability through a set of clear, rule-based decision paths. For instance, it identified that customers with month-to-month contracts and monthly charges above a certain threshold were significantly more likely to churn. Such transparent rules are highly beneficial for managerial teams in designing targeted retention campaigns and understanding the factors driving customer exits.

However, while the CHAID model offered interpretability, its predictive power was slightly lower compared to the **Logistic Regression model**. The Logistic Regression model achieved an **overall accuracy of 82.3%**, with strong performance across key evaluation metrics such as **ROC-AUC (0.87)** and **Precision-Recall balance**. The probabilistic nature of Logistic Regression allowed the estimation of churn likelihood for each customer, enabling prioritization of at-risk individuals for intervention.

The comparative analysis between both models showed that **Logistic Regression offers an optimal trade-off between interpretability and performance**, making it suitable for both technical deployment and strategic decision-making. It provides not only accurate churn predictions but also measurable coefficients that quantify the impact of each predictor variable.

Overall, the results demonstrate that data-driven churn prediction can effectively support proactive business decisions. By integrating model insights with customer relationship management (CRM) systems, telecom companies can identify early warning signs of dissatisfaction, optimize pricing strategies, and personalize retention offers. The findings underscore the practical value of combining **inferential statistics** with **machine learning** to develop a robust, deployable churn prediction solution that drives measurable business outcomes.


Customer Churn Prediction App
This app predicts whether a telecom customer is likely to churn.
Gender
Male
Senior Citizen
1
Partner
Yes
Dependents
Yes
Tenure (months)
12
Phone Service
No
Multiple Lines
No
Internet Service
No
Online Security
No internet service
Online Backup
No
Device Protection
No
Tech Support
No
Streaming TV
No internet service
Streaming Movies
No internet service
Contract
One year
Paperless Billing
No
Payment Method
Mailed check
Monthly Charges
10.00
Total Charges
200.00
Predict Churn
This customer is not likely to churn.

12. CONCLUSION

This project successfully demonstrates a **comprehensive end-to-end data science workflow** for predicting customer churn in the telecommunications sector. Starting from raw data, the process encompassed all major stages of a modern analytics pipeline — including **data cleaning, preprocessing, exploratory analysis, model development, evaluation, and deployment**. The integration of statistical and machine learning approaches ensured both interpretability and predictive precision, reflecting a balanced and practical solution to a real-world business challenge.

The analysis identified that **customer tenure, monthly charges, and contract type** are key predictors of churn. Customers on **short-term or month-to-month contracts** and those paying **higher monthly fees** were more likely to discontinue services. Similarly, **senior citizens** and users with **electronic check payments** showed slightly higher churn tendencies. These insights provide valuable guidance for designing targeted **retention and loyalty strategies**.

Two models — **CHAID Decision Tree** and **Logistic Regression** — were developed and compared. The CHAID model offered interpretable, rule-based insights into churn behavior, while the Logistic Regression model achieved superior predictive accuracy (**82.3%**) and a strong **ROC-AUC score**, making it the preferred choice for deployment. This comparative analysis highlights the importance of combining **statistical inference** with **machine learning** to balance interpretability and performance.

The final **Logistic Regression model** was deployed as an interactive **Streamlit web application**, allowing real-time churn prediction through a simple, user-friendly interface. This deployment bridges the gap between analytical modeling and managerial application, transforming the project into a **practical decision-support tool** for business users.

From a business perspective, the system enables telecom providers to proactively identify high-risk customers and implement **personalized retention strategies**, ultimately reducing acquisition costs and improving long-term customer loyalty.

Future Scope

Future enhancements may include:

- Integrating **ensemble techniques** (Random Forest, Gradient Boosting, XGBoost) to boost prediction robustness.
- Employing **explainable AI tools** like **SHAP** or **LIME** for greater model transparency.
- Automating **model retraining pipelines** to adapt to evolving customer behavior.
- Incorporating **sentiment analysis** and feedback data for deeper behavioral insights.

In summary, this project not only delivers a technically sound churn prediction model but also demonstrates how **AI-driven analytics** can enhance business decision-making and customer relationship management. With continued improvement and scalability, this framework can serve as a **core analytical component** for customer retention strategies in the telecom industry and beyond.

13. REFERENCES

1. Kaggle – *Telco Customer Churn Dataset.* Retrieved from https://www.kaggle.com/blastchar/telco-customer-churn
2. Scikit-learn Documentation. *Machine Learning in Python.* https://scikit-learn.org
3. Streamlit Official Documentation. *Build Data Apps for Machine Learning and Data Science.* https://docs.streamlit.io
4. Pandas, NumPy, and Matplotlib Libraries. *Python Data Analysis and Visualization Tools.*
5. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd Edition). Morgan Kaufmann Publishers.
6. SRM University – *Department of Computational Intelligence, Case Study Guidelines (2025).*
7. Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann.
8. Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* O'Reilly Media.
9. Brownlee, J. (2020). *Machine Learning Mastery with Python: Understand, Design, and Build Predictive Models.* Machine Learning Mastery.
10. IBM Analytics (2023). *Customer Churn Prediction Using Machine Learning.* IBM Cloud Documentation.
11. Xu, Y., Li, Q., & Wang, S. (2021). *Predicting Customer Churn in the Telecom Industry Using Machine Learning Techniques. Journal of Information and Computational Science*, 18(4), 245–252.
12. Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th Edition). Pearson Education.
13. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition). Springer.
14. Google Developers (2024). *Machine Learning Crash Course: Model Evaluation and Validation.* https://developers.google.com/machine-learning
15. Towards Data Science. (2022). *Understanding Customer Churn Prediction in Telecom Using Logistic Regression and Decision Trees.* Retrieved from https://towardsdatascience.com