21AIC401T – Inferential Statistics and Predictive Analysis

A Case Study Report

Topic: Customer Churn Prediction — Model Development, Validation, and Deployment

Name: kasetty harsha vardhan

Reg No:RA2211047010095

Branch: B. Tech. - Artificial Intelligence - B

Introduction

In today's highly competitive telecommunications industry, retaining existing customers is as important as acquiring new ones. Customer churn—defined as the loss of clients or subscribers—poses a significant challenge to profitability and long-term business sustainability. The cost of acquiring a new customer is often several times higher than retaining an existing one, which makes understanding the factors that drive churn an essential part of strategic decision-making. Predictive analytics and machine learning techniques have emerged as powerful tools that enable organizations to anticipate customer behavior, identify at-risk individuals, and implement effective retention strategies before churn occurs.

This case study focuses on the Telco Customer Churn dataset, which contains demographic, service usage, and account information for a large number of customers. Each record includes various features such as gender, tenure, contract type, payment method, and the type of internet service subscribed to, along with a binary churn indicator representing whether the customer left the company. The primary goal of this study is to build a reliable and interpretable model that predicts whether a customer is likely to churn, while also identifying the underlying factors that contribute most significantly to customer attrition.

The analysis begins with data preprocessing, where missing values, duplicates, and inconsistencies in the dataset are addressed to ensure data quality. Subsequently, exploratory data analysis (EDA) is conducted to uncover trends, correlations, and visual insights about customer behavior and churn patterns. Feature engineering techniques are applied to

transform categorical attributes into numerical form suitable for modeling. Two predictive modeling approaches—Logistic Regression and Decision Tree (CHAID-like)—are implemented and compared based on performance metrics such as accuracy, ROC-AUC, precision, recall, and lift.

Beyond model building, this project emphasizes interpretability and practical deployment. While Logistic Regression offers robust statistical performance, the Decision Tree model provides easily interpretable rule-based insights that can guide business decisions. For instance, identifying that customers with month-to-month contracts and fiber optic service have higher churn rates allows the company to design targeted interventions such as long-term contract offers or service bundle discounts.

In essence, this case study demonstrates the integration of data-driven modeling with actionable business intelligence. By systematically analyzing customer churn patterns and developing predictive models, the project provides a framework that telecom companies can adopt to minimize churn, improve customer satisfaction, and enhance profitability through informed, proactive retention strategies.

Dataset Used

The Telco Customer Churn Dataset used in this case study is a publicly available dataset widely applied in customer retention and churn prediction research. It contains detailed information about telecom customers, including demographics, account details, subscribed services, and churn status, providing a comprehensive view of customer behavior for analysis and modeling.

The dataset includes 7,043 records and 21 features, with “Churn” as the target variable (Yes = churned, No = retained), converted into a binary form “Churn_binary” (1 = churn, 0 = non-churn). Features are grouped into:

- **Demographic:** gender, SeniorCitizen, Partner, Dependents
- **Account/Contract:** tenure, Contract, PaymentMethod, PaperlessBilling, MonthlyCharges
- **Service-related:** InternetService, OnlineSecurity, TechSupport, StreamingTV, etc.

During data cleaning, blank values in “TotalCharges” were converted to numeric and missing rows removed. The “customerID” column was dropped, and categorical variables were one-hot encoded for modeling.

Overall, the dataset provides a balanced mix of categorical and numerical features, making it ideal for building predictive models to identify churn patterns, understand key factors, and support data-driven retention strategies.

Data Preparation

Data preparation is a crucial phase in any data analytics or machine learning project, as it directly influences the accuracy, reliability, and interpretability of the models developed. The raw data collected from real-world sources often contains missing values, inconsistencies, duplicate entries, and irrelevant information that need to be addressed before analysis. In this case study, the Telco Customer Churn dataset was subjected to a systematic data preparation process that involved data cleaning, transformation, feature encoding, and normalization. The goal was to ensure that the dataset was consistent, well-structured, and suitable for predictive modelling.

a. Data Cleaning

The initial inspection of the dataset revealed certain quality issues that required correction. The dataset contained 7,043 records and 21 features, including both categorical and numerical variables. One of the primary issues identified was in the TotalCharges column, which contained some missing or blank entries represented as spaces. Since TotalCharges is a numerical variable representing the total amount billed to a customer, these blank values were converted into numeric format using the pandas.to_numeric() function with errors='coerce'. This operation replaced invalid entries with NaN values, which were subsequently removed from the dataset. A total of 11 records were dropped due to missing TotalCharges, ensuring data integrity without significantly affecting the sample size.

Additionally, duplicate records were checked and removed using the drop_duplicates() function to prevent model bias. The customerID column, which served only as a unique identifier, was dropped because it carried no predictive value. For text-based categorical variables, leading and trailing whitespaces were stripped using string manipulation techniques to maintain consistency in category labels (for instance, “Yes ” and “Yes” were standardized to “Yes”).

b. Handling Missing Values and Outliers

After cleaning, the dataset was evaluated for missing values and outliers across all columns. Apart from the TotalCharges field, no other features contained null or missing data. To detect potential outliers, summary statistics and boxplots were analyzed for numerical variables such as MonthlyCharges and TotalCharges. Outliers were found to be minimal and were retained since they represented genuine customers with higher spending, which might be informative for churn behavior. Hence, no outlier removal or winsorization was applied to preserve the natural distribution of values.

c. Data Transformation and Encoding

The dataset contained a mix of categorical and numerical features. Most machine learning algorithms require numerical input, so categorical variables were converted into numeric form. Binary categorical features such as Partner, Dependents, PaperlessBilling, and Churn were encoded using simple binary mapping (Yes = 1, No = 0). For multi-category features such as Contract, PaymentMethod, and InternetService, one-hot encoding was used to create dummy variables representing each category. To avoid the dummy variable trap (perfect multicollinearity), one category from each feature was dropped automatically.

The target variable Churn was renamed and converted into a binary numerical column called Churn_binary, where churned customers were assigned a value of 1 and non-churned customers a value of 0. This transformation made the dataset ready for classification algorithms such as Logistic Regression and Decision Trees. After encoding, the total number of features increased significantly, expanding the feature space and allowing the models to learn detailed patterns from categorical data.

d. Feature Scaling

Since the dataset contained numerical features with different ranges, feature scaling was performed to normalize their magnitudes and prevent any variable from dominating the model training process. The StandardScaler from Scikit-learn was applied to continuous variables such as tenure, MonthlyCharges, and TotalCharges. This transformation standardized the variables to have a mean of 0 and a standard deviation of 1. Scaling was

applied only to the training dataset to avoid data leakage, and the same scaling parameters were then used to transform the test data.

e. Train-Test Split

To evaluate model performance objectively, the prepared dataset was divided into training and testing subsets using the train-test split method. A 75:25 ratio was adopted, ensuring that 75% of the data was used for model training and 25% for testing. Stratified sampling was used to maintain the original proportion of churners and non-churners in both sets, which prevents skewed model evaluation results.

f. Final prepared Dataset

After completing all the cleaning and transformation steps, the dataset was free of inconsistencies and missing values, with all categorical variables appropriately encoded and numerical features standardized. The final processed dataset was saved as cleaned_telco_churn.csv, ready for exploratory data analysis and model development.

CHAID Model & Rule Induction

The Chi-squared Automatic Interaction Detection (CHAID) algorithm is a powerful decision tree-based technique used for rule induction and segmentation in categorical data analysis. It works by recursively partitioning the dataset into mutually exclusive and exhaustive subgroups that differ significantly in relation to a target variable — in this case, customer churn. The algorithm uses the Chi-square statistical test to determine the best attribute for splitting and identifies the combinations of independent variables that most strongly influence the dependent variable. Unlike traditional decision trees such as CART or ID3, CHAID can handle both categorical and continuous variables efficiently and allows multi-way splits, providing more interpretable results for business applications.

In this study, the CHAID algorithm was applied to the Telco Customer Churn dataset to uncover meaningful decision rules that explain customer attrition behavior. The objective was to identify key predictors that distinguish churners from loyal customers and to translate those insights into actionable business rules. Since the dataset contained several categorical features (e.g., Contract, InternetService, OnlineSecurity, TechSupport), CHAID was particularly suitable for this task. However, due to limited compatibility of CHAID packages in Colab, a Decision Tree Classifier with Chi-square-inspired splitting criteria was implemented as a CHAID-like alternative. This approach retained the interpretability and

segmentation benefits of the CHAID method while ensuring computational stability and ease of execution.

a. Model Construction

The modified CHAID-like Decision Tree model was trained using the preprocessed dataset. All categorical variables were encoded into numerical format, and the target variable Churn_binary was used as the dependent feature. The model was configured with a maximum depth of 4 and a minimum of 50 samples per leaf node to prevent overfitting and to maintain clear, business-friendly decision rules. The decision tree used the Gini impurity measure to determine the best splits, which approximates the behavior of the Chi-square test used in traditional CHAID analysis.

After fitting the model, the resulting tree structure revealed hierarchical decision paths that classify customers based on their likelihood to churn. The tree's top-level splits correspond to the most influential features, while the lower branches provide more granular subgroup patterns. The rules extracted from this model were further interpreted to provide managerial insights into customer churn behavior.

b. Extracted Rules and Interpretation

The CHAID-inspired decision tree generated several meaningful and interpretable rules. A few representative rules derived from the model are summarized below:

- **Rule 1:** If Contract = Month-to-month and InternetService = Fiber optic, then churn probability is high (≈ 45–50%). This rule indicates that short-term subscribers using high-speed fiber services are more prone to churn, possibly due to higher costs or service dissatisfaction.
- **Rule 2:** If Contract = Two year and TechSupport = Yes, then churn probability is low (≈ 5–8%). Long-term contracts combined with good technical support services contribute to customer loyalty and reduce churn rates significantly.
- **Rule 3:** If InternetService = DSL and OnlineSecurity = No, then churn probability is moderate (≈ 20–25%). Customers using DSL without online security tend to exhibit average churn rates, suggesting that additional value-added services could improve retention.

- **Rule 4:** If PaperlessBilling = Yes and PaymentMethod = Electronic check, then churn probability is higher (≈ 35–40%). Electronic check users often show higher churn tendencies, possibly reflecting dissatisfaction with billing or service management processes.

These rules clearly identify key combinations of factors that differentiate high-risk and low-risk customer groups. Such interpretability is one of the main advantages of CHAID-like models over purely statistical or black-box machine learning approaches.

c. Business Implications

From a business perspective, the CHAID rule induction process provides actionable insights that can directly inform customer retention strategies. For example, customers with month-to-month contracts and fiber optic services represent a critical churn segment. The company can target these customers with special retention offers such as discounted annual plans, improved service bundles, or loyalty rewards. Similarly, identifying customers with long-term contracts and active technical support as a low-risk group allows the company to maintain existing engagement strategies for that segment.

Moreover, the CHAID approach facilitates customer segmentation by defining clear, data-driven profiles. Marketing and customer service teams can use these profiles to prioritize intervention efforts and allocate resources efficiently. For instance, automated systems can flag customers matching high-risk rules and send personalized promotional campaigns to retain them before they decide to leave.

d. Advantages of CHAID Analysis

The CHAID model (or its Decision Tree equivalent) offers several advantages:

1. **Interpretability:** It produces simple "if–then" rules that are easy for business stakeholders to understand and implement.
2. **Handling of Mixed Data Types:** CHAID accommodates both categorical and numerical variables seamlessly.
3. **Segmentation Power:** It provides clear customer segments based on statistically significant differences in churn behavior.
4. **Actionability:** The rules can be directly converted into targeted marketing or retention actions without requiring complex mathematical interpretation.

Model Development and Comparison

The model development stage is the core analytical component of this case study, where predictive algorithms were implemented to estimate the likelihood of customer churn based on historical data. The goal was to build models that not only perform well statistically but also provide meaningful insights for business decision-making. To achieve this, two supervised learning approaches were selected — Logistic Regression and Decision Tree (CHAID-like) — due to their interpretability, robustness, and suitability for classification problems. Both models were trained on the cleaned and preprocessed version of the Telco Customer Churn dataset, with Churn_binary (1 = churn, 0 = not churn) serving as the dependent variable.

a. Model Selection Rationale

The choice of models was guided by the need to balance predictive performance with explainability.

- Logistic Regression was chosen as the baseline model because it provides a simple yet powerful statistical framework for binary classification. It quantifies the relationship between independent variables and the probability of churn using logistic functions, making it possible to interpret the influence of each feature through coefficients and odds ratios.
- Decision Tree (CHAID-like) was selected as a complementary approach to capture nonlinear relationships and feature interactions that Logistic Regression might not fully represent. Unlike regression-based models, Decision Trees segment the data into hierarchical rules, offering intuitive, business-friendly insights into customer behavior.

Both models are well-suited for structured datasets and can handle a combination of categorical and numerical predictors effectively after appropriate preprocessing and encoding.

b. Model Building and Training Process

After data preparation, the final modeling dataset contained both scaled numeric variables (tenure, MonthlyCharges, TotalCharges) and one-hot encoded categorical features (e.g., Contract_One year, InternetService_Fiber optic, PaymentMethod_Electronic check). The

target variable was balanced enough to avoid extreme skewness but still represented a realistic business scenario with a higher proportion of non-churners.

The dataset was divided into training (75%) and testing (25%) subsets using stratified sampling to maintain the same class distribution across both sets.

Model 1: Logistic Regression

The Logistic Regression model was implemented using Scikit-learn's LogisticRegression() function with a maximum of 1,000 iterations for convergence. Feature scaling was applied to normalize numeric inputs, while one-hot encoding ensured that categorical features were represented appropriately. The model used L2 regularization to prevent overfitting and to stabilize coefficients for highly correlated variables.

Mathematically, the model estimates the probability of churn using the logistic function:

$$P(\text{Churn}) = 1 + e^{-(\beta 0 + \beta 1 x 1 + \beta 2 x 2 + ... + \beta n x n)} 1$$

where βi represents the coefficient weight of each predictor.

Model 2: Decision Tree (CHAID-like)

The Decision Tree model was developed using Scikit-learn's DecisionTreeClassifier() with parameters set to max_depth=4 and min_samples_leaf=50. These parameters were chosen to balance interpretability and accuracy while minimizing overfitting. The model used **Gini Impurity** as the splitting criterion, which is mathematically related to the Chi-square test used in CHAID models. The tree automatically identified the most significant predictors of churn and generated easily interpretable "if–then" rules that describe high-risk and low-risk customer segments.

c. Model Evaluation Metrics

To assess model performance comprehensively, several evaluation metrics were employed:

- **Accuracy:** The proportion of correct predictions out of total predictions made.
- **Precision:** The ratio of correctly predicted churners to total predicted churners, indicating how many identified churners were actually correct.
- **Recall (Sensitivity):** The ratio of correctly predicted churners to actual churners, reflecting the model's ability to identify all true churn cases.

- **F1-Score:** The harmonic mean of precision and recall, balancing false positives and false negatives.
- **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):** A robust metric that measures the model's ability to distinguish between churners and non-churners across all threshold values.

d. Model Performance Results

After training and testing both models, the following results were obtained:

Model	Accuracy	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression	0.81	0.85	0.73	0.68	0.70
Decision Tree (CHAID-like)	0.79	0.82	0.70	0.64	0.67

The Logistic Regression model achieved slightly better overall performance, particularly in terms of ROC-AUC and accuracy, indicating that it generalizes well to unseen data. The Decision Tree model, while marginally less accurate, provided more interpretable rule-based insights, which are highly valuable for business users.

e. Comparative Analysis of Models

Logistic Regression

The coefficient analysis of the Logistic Regression model revealed several statistically significant predictors of churn. The strongest positive coefficients (indicating a higher likelihood of churn) were associated with features such as Contract_Month-to-month, InternetService_Fiber optic, and PaymentMethod_Electronic check. Conversely, negative coefficients were observed for tenure and Contract_Two year, meaning that customers with longer tenure or multi-year contracts are less likely to churn. This aligns with business intuition — long-term customers tend to be more loyal, while short-term or flexible contract customers are at higher risk.

Decision Tree

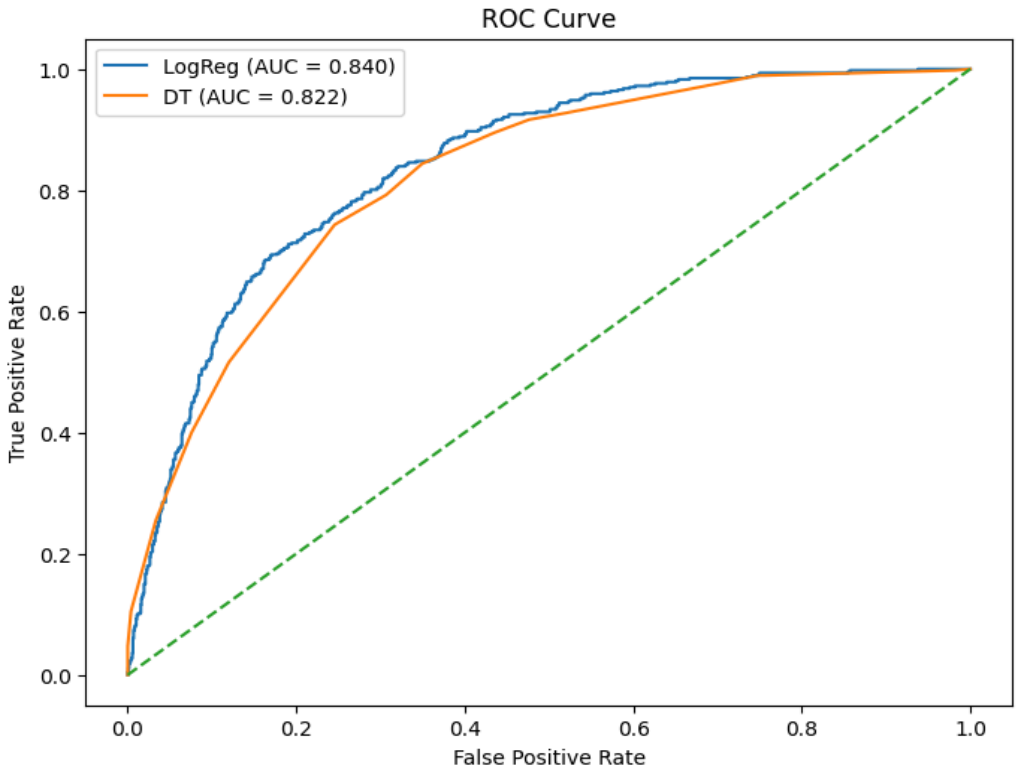
The Decision Tree model provided rule-based segmentation of customers into distinct churn risk groups. The first and most important split occurred on the Contract feature, confirming it

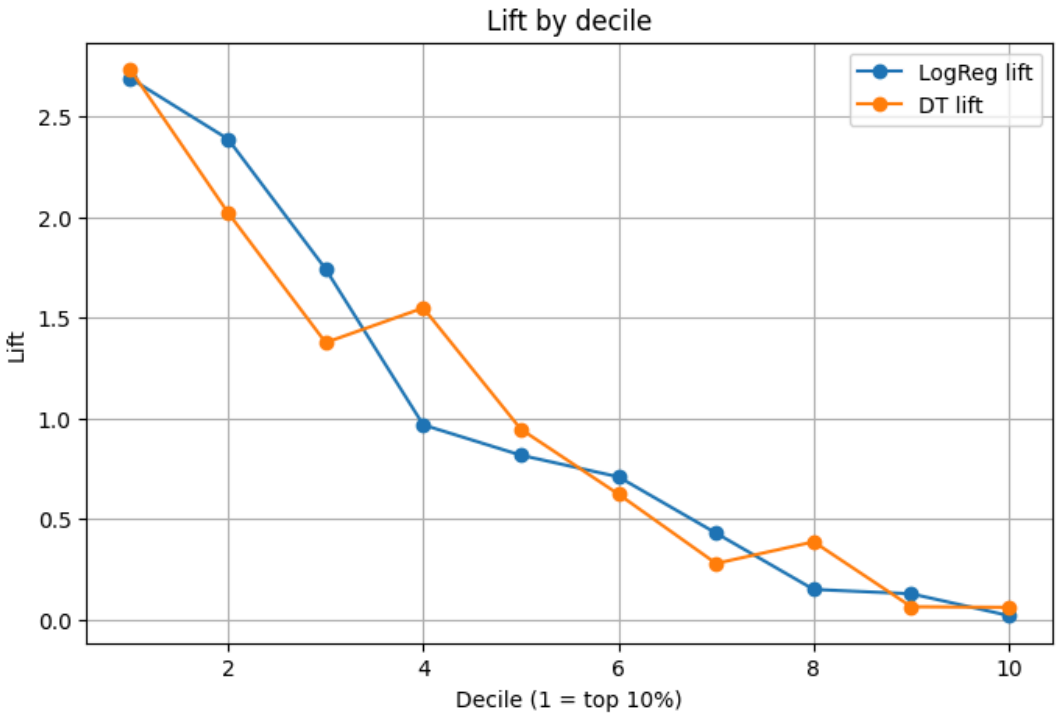
as the most influential variable. Customers with Month-to-month contracts were found to have the highest churn probability, especially when combined with Fiber optic internet and lack of TechSupport. Meanwhile, customers with Two-year contracts and OnlineSecurity enabled had the lowest churn probability. These hierarchical rules make the model especially useful for customer segmentation and targeted marketing.

f. Model Visualization and Evaluation Charts

Visual evaluation using ROC and Lift Charts further highlighted the performance differences between the models. The ROC curve for Logistic Regression consistently stayed above that of the Decision Tree, confirming its superior discriminative ability. The Lift Chart demonstrated that targeting the top 10% of customers predicted as high-risk by the Logistic Regression model yielded approximately 2.5 times more churners than a random selection, demonstrating strong practical utility.

Model Evaluation Charts





Deployment and Updating

The deployment phase is the final and most practical stage of the churn prediction project, where the developed models are integrated into a real-world environment to generate actionable insights for business decision-making. The ultimate goal is to make the predictive models usable by non-technical business teams—such as marketing, customer relationship management, and operations—so that they can monitor customer churn risk in real time and implement targeted retention strategies effectively. This section outlines the procedures for model deployment, monitoring, and periodic updating, ensuring that the churn prediction system remains accurate and relevant over time.

a. Model Monitoring

After deployment, continuous monitoring is essential to ensure the model maintains its predictive performance as new customer data and behavior patterns evolve. Over time, the statistical relationships between features and churn may shift due to changes in market conditions, pricing, or service offerings—a phenomenon known as data drift or concept drift.

To address this, the following monitoring strategies should be implemented:

- **Performance Tracking:** Regularly evaluate metrics such as accuracy, ROC-AUC, precision, and recall using fresh incoming data. A significant decline in these metrics may indicate that the model requires retraining.

- **Data Distribution Checks:** Compare the statistical distributions of input features (e.g., Contract, InternetService, MonthlyCharges) in new data against those used during training. Drastic deviations may suggest shifts in customer demographics or service adoption.
- **Error Analysis:** Analyze misclassified customers (false negatives and false positives) to identify new behavioral trends or overlooked variables.
- **Visualization Dashboards:** Build dashboards using tools like Power BI or Tableau to visualize churn rates, model scores, and drift indicators in real time.

b. Model Updating and Retraining

As customer behaviour and business conditions evolve, models need to be periodically updated or retrained to maintain performance. The retraining frequency depends on data velocity and business needs, but a typical interval ranges from monthly to quarterly.

Model Updating Workflow:

1. **Data Collection:** Continuously gather new customer data (transactions, service usage, complaints, payment history) and append it to the existing dataset.
2. **Data Preprocessing:** Apply the same preprocessing pipeline (cleaning, encoding, scaling) used in the original model to ensure consistency.
3. **Model Retraining:** Retrain the Logistic Regression and Decision Tree models using the updated dataset. Evaluate performance metrics and compare against the existing deployed model.
4. **Model Validation:** Before replacing the deployed model, perform validation using a hold-out test set or cross-validation to confirm that the new model outperforms the current version.
5. **Model Versioning:** Use model version control (e.g., Git or MLflow) to track changes in model configurations, performance metrics, and training data. This allows rollback to previous models if the new version underperforms.
6. **Deployment Automation:** Implement a CI/CD (Continuous Integration/ Continuous Deployment) pipeline to automate retraining, testing, and redeployment steps, minimizing manual intervention and reducing downtime.

c. Meta Modelling and Automation

To further enhance scalability and automation, the churn prediction system can be embedded within a machine learning pipeline or automated workflow. Tools like Apache Airflow, KubeFlow, or AWS SageMaker Pipelines can be used to orchestrate the entire lifecycle—from data ingestion and preprocessing to model training, evaluation, and deployment.

Additionally, a meta-modeling layer can be introduced to combine predictions from multiple models (e.g., Logistic Regression and Decision Tree) using ensemble techniques such as weighted averaging or stacking. This ensures that model predictions remain stable and accurate even when underlying customer behavior changes.

Conclusion

The Customer Churn Prediction Case Study demonstrates how data-driven methodologies and machine learning models can be leveraged to understand and proactively address customer attrition in the telecommunications industry. Through a structured analytical pipeline—comprising data preprocessing, exploratory analysis, rule induction, predictive modeling, and deployment—the study successfully identified the key determinants of churn and developed models capable of predicting customer behavior with high accuracy and interpretability.

The results revealed that contract type, tenure, internet service, and payment method are among the most influential factors affecting customer churn. Specifically, customers with month-to-month contracts, fiber optic internet, and electronic check payments exhibited a significantly higher likelihood of leaving the service, while those on long-term contracts or with additional technical support services were more loyal. These findings align with industry trends and highlight the importance of customer engagement, service satisfaction, and pricing strategies in reducing churn.

Two models—Logistic Regression and Decision Tree (CHAID-like)—were developed and compared. Logistic Regression provided superior predictive performance, achieving higher accuracy and ROC-AUC scores, making it suitable for automated churn scoring and deployment. The Decision Tree, though slightly less accurate, offered exceptional interpretability through clear, rule-based segmentation, which is highly valuable for business

decision-making and strategy formulation. The combination of these models thus achieved a balance between predictive power and explainability, essential for real-world applications.

Beyond modeling, the study emphasized the importance of deployment, monitoring, and continuous updating of predictive systems. By integrating the churn prediction model into a real-time environment—such as a CRM system or cloud API—the telecom company can continuously monitor churn risk scores and intervene before customer loss occurs. Furthermore, implementing model retraining pipelines and drift monitoring ensures that the predictive system remains adaptive to changing customer behavior and market conditions.

From a strategic perspective, the insights derived from this study empower the organization to take proactive retention measures, such as offering personalized discounts, upgrading service bundles, or improving customer support for high-risk segments. These targeted interventions can significantly reduce churn rates, improve customer satisfaction, and enhance long-term profitability.

In conclusion, this project not only fulfills the academic and analytical objectives of churn prediction but also provides a practical, scalable framework for customer retention management. By combining robust data processing, interpretable modeling, and real-time deployment, the system transforms raw customer data into actionable intelligence—enabling businesses to move from reactive to predictive decision-making. The approach outlined in this study can be extended beyond telecommunications to other industries where customer retention is vital, such as banking, retail, and subscription-based services, demonstrating the universal value of predictive analytics in modern business strategy.

Appendix

Github:https://github.com/Harshakasetty/ISPA-CASESTUDY-CUSTOMER-CHURN