SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

GREATER NOIDA, UTTAR PRADESH

2025-2026

B.tech CSE (Data Science)

MACHINE LEARNING (Course Code: R1UC525B)

A Case Study Report

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

Submitted By:

Team Size: 2

- Student 1: VAIBHAV DWIVEDI (23SCSE1420093)
- Student 2: SHREYA YADAV (23SCSE1420150)

Submitted To:

Prof. Akhilesh Kumar Singh

Date : 28/12/2025

INDEX

Customer Churn Prediction Using Machine Learning

ABSTRACT

Customer churn prediction is an important problem for service-oriented industries, as losing existing customers directly affects revenue and long-term business growth. With the increasing availability of large-scale customer data, machine learning techniques provide effective solutions for identifying customers who are likely to discontinue a service.

This case study presents a machine learning-based approach for predicting customer churn using a publicly available telecom dataset. The dataset consists of customer demographic information, service usage details, and billing attributes. The data was cleaned, pre-processed, and divided into training, validation, and testing sets to ensure reliable model evaluation. Two supervised classification algorithms, namely Logistic Regression and Random Forest, were implemented to perform churn prediction.

The performance of both models was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve. Experimental results show that the Random Forest model outperforms Logistic Regression across most evaluation metrics, indicating its ability to capture complex patterns in customer behavior.

The findings of this study demonstrate that machine learning models can effectively support customer retention strategies by enabling early identification of high-risk customers. This work highlights the practical application of predictive analytics in business decision-making and provides scope for further improvements using advanced models and real-time data.

1. INTRODUCTION & MOTIVATION

In recent years, rapid digitalization and technological advancement have transformed the way organizations interact with their customers. Service-based industries such as telecommunications, banking, insurance, and subscription-based platforms operate in highly competitive environments where customers have multiple alternatives available. In such scenarios, retaining existing customers has become more challenging yet more important than ever. Customer churn, defined as the phenomenon in which customers discontinue a service or switch to a competitor, poses a serious threat to organizational profitability and sustainability.

Customer acquisition typically involves high marketing and operational costs, whereas retaining an existing customer is comparatively less expensive and more beneficial in the long term. Studies have shown that even a small reduction in churn rate can lead to a significant increase in overall revenue. Therefore, identifying customers who are likely to churn at an early stage is a critical business requirement. However, traditional approaches to churn identification rely on manual analysis, rule-based systems, or subjective judgment, which are often inefficient, inaccurate, and unsuitable for large-scale data.

With the growth of digital platforms, organizations now collect vast amounts of customer-related data, including demographic details, service usage patterns, billing information, and interaction history. Analyzing such high-dimensional data manually is impractical. This is where Machine Learning (ML) plays a crucial role. Machine learning techniques enable automated data analysis, pattern recognition, and predictive modeling, allowing organizations to extract meaningful insights from historical data.

Machine learning-based churn prediction models learn from past customer behavior and identify patterns associated with churn. These models can predict the probability of a customer leaving the service before it actually occurs. Such predictive capabilities allow companies to take proactive retention measures such as personalized offers, loyalty programs, improved customer support, and targeted marketing campaigns. as a result, businesses can enhance customer satisfaction, reduce revenue loss, and gain a competitive advantage.

The motivation behind this case study is to explore the application of supervised machine learning algorithms for predicting customer churn using real-world data. This study aims to provide hands-on experience in implementing an end-to-end machine learning pipeline, including data pre-processing, feature encoding, model training, and performance evaluation. Additionally, the project seeks to compare the effectiveness of different classification algorithms and analyze their strengths and limitations in handling customer churn prediction tasks.

By conducting this case study, the practical relevance of machine learning in business analytics is demonstrated. The work emphasizes how data-driven decision-making can help organizations shift from reactive to proactive customer retention strategies. Furthermore, this study serves as a foundation for future enhancements such as the use of advanced models, real-time data integration, and deployment of churn prediction systems in real-world applications.

2. PROBLEM STATEMENT & OBJECTIVES

2.1 Problem Statement

Customer churn is one of the most significant challenges faced by service-based industries, particularly in highly competitive sectors such as telecommunications. Customers today have easy access to multiple service providers offering similar services at comparable prices. As a result, even minor dissatisfaction can lead to customers switching to competitors. This frequent switching behavior makes it difficult for organizations to maintain stable revenue and long-term customer relationships.

The primary problem addressed in this case study is the lack of an efficient and accurate mechanism to identify customers who are likely to churn in the near future. Traditional churn detection methods are often reactive in nature, identifying churn only after the customer has already left. Such approaches do not provide sufficient time for organizations to implement retention strategies and prevent customer loss.

Furthermore, customer data is typically large, complex, and multidimensional, consisting of demographic information, service usage patterns, billing details, and contractual attributes. Manual analysis of such data is impractical and prone to human error. Therefore, there is a need for an automated, data-driven approach that can analyze historical customer data and predict churn behavior effectively.

This project addresses the problem by developing a supervised machine learning-based classification system that predicts whether a customer will churn or not. By learning patterns from historical data, the proposed system aims to assist organizations in identifying high-risk customers in advance and enabling timely intervention.

2.2 Objectives

The main objective of this case study is to design, implement, and evaluate machine learning models for customer churn prediction. The specific objectives of the study are as follows:

1. To analyze customer data and understand the factors that influence customer churn behavior.
2. To perform data cleaning and pre-processing to improve data quality and prepare it for machine learning models.
3. To build supervised classification models for predicting customer churn using historical data.
4. To implement and compare multiple machine learning algorithms on the same dataset to ensure fair evaluation.
5. To evaluate model performance using standard classification metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve.
6. To identify the most effective model for churn prediction based on experimental results.
7. To demonstrate the practical application of machine learning in solving real-world business problems related to customer retention.

3. DATASET DESCRIPTION

The dataset used for this case study is the **Telco Customer Churn Dataset**, which is a widely used benchmark dataset for customer churn analysis in the telecommunications domain. The dataset is publicly available and was obtained from **Kaggle**, ensuring transparency, reproducibility, and ethical use of data for academic purposes. This dataset represents real-world customer information collected by a telecom service provider and is suitable for studying customer behavior and churn patterns.

3.1 Dataset Source and Nature

- **Dataset Name:** Telco Customer Churn Dataset
- **Source:** Kaggle (Public Dataset)
- **Domain:** Telecommunications
- **Data Type:** Tabular structured data
- **Learning Type:** Supervised learning
- **Target Variable:** Churn (Yes / No)

The dataset contains detailed records of customers along with their demographic characteristics, subscribed services, billing information, and contractual details. Each row in the dataset corresponds to an individual customer, and each column represents a specific attribute related to that customer.

3.2 Dataset Size and Composition

- **Total Records:** 7043 customer instances
- **Total Features:** 21 attributes (including target variable)
- **Input Features:** 20
- **Target Feature:** 1 (Churn)

The dataset size is sufficient to train and evaluate machine learning models effectively and satisfies the course requirement of having more than 300–500 records with multiple input features.

3.3 Feature Categories

The features in the dataset can be broadly classified into the following categories:

1. Customer Demographic Information

These attributes describe the personal characteristics of customers.

- **gender:** Indicates whether the customer is male or female
- **SeniorCitizen:** Binary variable indicating whether the customer is a senior citizen
- **Partner:** Indicates whether the customer has a partner
- **Dependents:** Indicates whether the customer has dependents

Demographic features help in understanding how different customer groups behave and how churn varies across age and family status.

2. Account and Subscription Information

These attributes describe the customer's account history and subscription details.

- **tenure:** Number of months the customer has stayed with the company
- **Contract:** Type of contract (month-to-month, one year, two year)
- **PaperlessBilling:** Indicates whether the customer uses paperless billing
- **PaymentMethod:** Method used for payment

Tenure is one of the most important predictors of churn, as customers with shorter tenure are generally more likely to leave.

3. Service Usage Information

These attributes indicate the services subscribed by the customer.

- **PhoneService**
- **MultipleLines**
- **InternetService**
- **OnlineSecurity**
- **OnlineBackup**
- **DeviceProtection**
- **TechSupport**
- **StreamingTV**
- **StreamingMovies**

Service usage patterns provide insights into customer engagement levels and satisfaction.

4. Billing and Financial Information

These attributes describe the customer's financial interactions with the service provider.

- **MonthlyCharges:** Amount charged to the customer per month
- **TotalCharges:** Total amount billed to the customer over the tenure period

Billing-related attributes are crucial for identifying customers who may be dissatisfied due to high costs.

3.4 Target Variable Description

- **Churn:** Indicates whether a customer has discontinued the service.
    - **Yes:** Customer has churned

    - **No:** Customer is still active

This variable is the output label for the machine learning classification task.

3.5 Data Quality and Characteristics

During initial data exploration, the dataset exhibited the following characteristics:

- Presence of missing or inconsistent values in the TotalCharges attribute
- Categorical variables represented as text labels
- Mixed data types (numerical and categorical)
- Slight class imbalance between churned and non-churned customers

These characteristics necessitated thorough data pre-processing before model training.

3.6 Suitability of Dataset for the Study

The Telco Customer Churn Dataset is well-suited for this case study due to the following reasons:

- Real-world relevance to business problems
- Sufficient size and feature richness
- Presence of both numerical and categorical attributes
- Clear target variable for supervised learning
- Compatibility with multiple machine learning algorithms

The dataset enables comprehensive analysis of customer behavior and supports meaningful comparison of classification models.

3.7 Ethical Considerations

The dataset is publicly available and anonymized, ensuring that no personally identifiable information is disclosed. The study adheres to ethical standards by using the data strictly for academic and research purposes.

4. DATA PRE-PROCESSING

Data pre-processing is a critical step in the machine learning pipeline, as the quality of input data directly influences the performance and reliability of predictive models. Real-world datasets often contain missing values, inconsistent data formats, irrelevant attributes, and categorical variables that cannot be directly processed by machine learning algorithms. Therefore, appropriate data cleaning and transformation steps were applied to prepare the dataset for effective model training and evaluation.

4.1 Removal of Irrelevant Attributes

The dataset initially contained a unique customer identifier attribute (customerID). While this attribute uniquely distinguishes customers, it does not provide any predictive information regarding churn behavior. Including such identifiers may introduce noise and negatively impact model performance. Therefore, the customerID column was removed from the dataset to ensure that only meaningful features were used for learning.

4.2 Handling Missing and Inconsistent Values

During exploratory analysis, it was observed that the TotalCharges attribute contained missing or invalid values. These missing values primarily occurred due to newly added customers with very low tenure, where total charges had not yet been accumulated.

To address this issue:

- The TotalCharges column was first converted from string format to numeric format.
- Missing values were then handled using mean imputation, where missing entries were replaced with the average value of the column.

Mean imputation was chosen as it preserves the overall distribution of the data and prevents loss of valuable records, which is particularly important for maintaining dataset size and balance.

4.3 Encoding of Categorical Variables

The dataset contained several categorical attributes such as gender, contract type, internet service, and payment method. Since most machine learning algorithms require numerical input, these categorical features needed to be converted into numerical form.

Label Encoding was applied to transform categorical variables into integer values. Each distinct category was assigned a unique numerical label. This approach was selected due to its simplicity and suitability for tree-based models such as Random Forest. Consistent encoding ensured that all categorical features were represented in a format compatible with the learning algorithms.

4.4 Feature–Target Separation

After cleaning and encoding the dataset, features and target variables were separated. All attributes except the Churn column were treated as input features, while the Churn attribute was designated as the target variable. This separation is essential for supervised learning, as it allows the model to learn the relationship between input variables and the output label.

4.5 Feature Scaling

The dataset contained numerical features with different value ranges, such as tenure, MonthlyCharges, and TotalCharges. Differences in scale can adversely affect certain machine learning algorithms, especially those based on distance or gradient optimization.

To address this issue, feature scaling was performed using **StandardScaler**, which standardizes features by removing the mean and scaling to unit variance. Feature scaling ensures that all numerical attributes contribute equally to the learning process and improves model convergence and stability.

4.6 Train–Validation–Test Split

To evaluate model performance fairly and prevent overfitting, the dataset was divided into three subsets:

- **Training Set (70%)**: Used to train the machine learning models
- **Validation Set (15%)**: Used for model tuning and performance monitoring
- **Testing Set (15%)**: Used for final performance evaluation

This three-way split ensures that the model is evaluated on unseen data and provides a reliable estimate of real-world performance.

4.7 Justification of Pre-processing Steps

Each pre-processing step was carefully selected to enhance data quality and model effectiveness:

- Removing irrelevant attributes reduced noise
- Handling missing values preserved dataset completeness
- Encoding categorical variables enabled algorithm compatibility
- Feature scaling improved learning efficiency
- Dataset splitting ensured unbiased evaluation

These pre-processing steps collectively ensured that the dataset was clean, consistent, and suitable for training robust machine learning models.

4.8 Outcome of Pre-processing

After applying all pre-processing techniques, the dataset was transformed into a structured and machine-readable format. The cleaned and scaled data provided a strong foundation for building and comparing machine learning models for customer churn prediction.

5.EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is an essential step in the machine learning workflow that helps in understanding the underlying structure, patterns, and relationships present in the dataset. Before building predictive models, it is important to analyze the data visually and statistically to gain insights into customer behavior and identify key factors influencing churn.

EDA was performed using statistical summaries and graphical visualizations to explore distributions, detect patterns, and observe relationships between features and the target variable.

5.1 Churn Distribution Analysis

The first step in EDA was to analyze the distribution of the target variable, Churn. This analysis helps in understanding whether the dataset is balanced or imbalanced.

The churn distribution revealed that:

- A larger proportion of customers did not churn.
- A smaller but significant portion of customers churned.

This slight class imbalance is common in real-world datasets and highlights the importance of using evaluation metrics beyond simple accuracy, such as recall and F1-score, to properly assess model performance.

5.2 Tenure vs Churn Analysis

Tenure represents the number of months a customer has remained with the company. An analysis of tenure against churn revealed a strong relationship between these two variables.

Key observations:

- Customers with shorter tenure exhibited a higher churn rate.
- Long-term customers were significantly less likely to churn.

This indicates that customer loyalty increases over time and that early-stage customers require greater attention to prevent churn. Tenure emerged as one of the most influential features in predicting churn.

5.3 Contract Type vs Churn

The relationship between contract type and churn was analyzed to understand how contractual commitments affect customer retention.

Findings include:

- Customers on month-to-month contracts had the highest churn rate.
- Customers with one-year or two-year contracts showed much lower churn rates.

This observation suggests that long-term contracts encourage customer retention and reduce churn probability. Contract type plays a critical role in churn prediction.

5.4 Monthly Charges vs Churn

MonthlyCharges reflect the recurring cost paid by customers for services. Analysis of this feature revealed noticeable trends:

- Customers with higher monthly charges showed a higher likelihood of churn.
- Lower monthly charges were associated with higher customer retention.

This pattern suggests that pricing strategies and perceived value for money strongly influence customer satisfaction and retention.

5.5 Internet Service and Churn

InternetService was another important feature analyzed during EDA.

Key insights:

- Customers using fiber optic internet services had higher churn rates compared to DSL users.
- Customers without internet services showed the lowest churn rates.

This indicates that service quality and pricing related to specific internet services may impact customer decisions to continue or discontinue services.

5.6 Payment Method Analysis

The relationship between payment methods and churn was also explored.

Observations:

- Customers using electronic check payment methods showed higher churn rates.
- Customers using automatic payment methods (credit card or bank transfer) were more likely to stay.

This insight suggests that convenience and payment automation may positively influence customer retention.

5.7 Correlation Analysis

A correlation analysis was performed on numerical features to understand their relationships with churn.

- Tenure showed a strong negative correlation with churn.
- MonthlyCharges showed a moderate positive correlation with churn.
- TotalCharges showed an indirect relationship influenced by tenure.

Correlation analysis helped identify the most impactful numerical features and guided model selection and feature importance interpretation.

5.8 Summary of EDA Findings

The exploratory data analysis provided several important insights:

- Customer churn is influenced by tenure, contract type, monthly charges, and service usage.
- Early-stage customers and those on flexible contracts are more likely to churn.
- Pricing and service quality play a crucial role in customer retention.

These insights justified the selection of relevant features and reinforced the need for machine learning models capable of capturing complex relationships in the data.

5.9 Role of EDA in Model Development

The insights gained from EDA informed the feature selection process and helped in understanding which variables were most influential in churn prediction. This understanding improved model interpretability and supported more accurate prediction results in later stages of the project.

6. MACHINE LEARNING MODELS & MODEL TRAINING

Machine learning models form the core of any predictive analytics system. The selection of appropriate algorithms and a well-defined training strategy are crucial for building reliable and accurate prediction systems. In this case study, two supervised machine learning classification models were selected and implemented to predict customer churn. These models were chosen to allow fair comparison between a simple, interpretable model and a more complex ensemble-based model.

6.1 Overview of Model Selection

The primary goal of this study was to compare the performance of different machine learning algorithms on the same dataset under identical conditions. Therefore, the following criteria were considered while selecting the models:

- Suitability for binary classification problems
- Ability to handle mixed numerical and categorical data
- Interpretability and practical relevance
- Widespread usage in industry and academia

Based on these considerations, **Logistic Regression** and **Random Forest Classifier** were selected for implementation.

6.2 Logistic Regression Model

6.2.1 Model Description

Logistic Regression is a widely used supervised learning algorithm for binary classification problems. Despite its name, it is a classification algorithm rather than a regression algorithm. Logistic Regression models the probability of a binary outcome using a logistic (sigmoid) function.

The hypothesis function of Logistic Regression is given by:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

where:

- $x_1, x_2, \ldots, x_n$are input features
- $\beta_0, \beta_1, \ldots, \beta_n$are model parameters

The output probability is mapped to a class label using a decision threshold, typically 0.5.

6.2.2 Why Logistic Regression Was Chosen

Logistic Regression was selected as:

- It provides a strong baseline for classification tasks
- It is simple, efficient, and interpretable
- It works well when the relationship between features and target is approximately linear

- It allows understanding the impact of individual features on churn prediction

6.2.3 Training Logistic Regression

The Logistic Regression model was trained using the training dataset after pre-processing and feature scaling. Feature scaling was essential for this model because it relies on gradient-based optimization techniques.

Key training details:

- Optimization method: Gradient Descent
- Regularization: Default L2 regularization
- Input: Scaled training features
- Output: Binary churn prediction

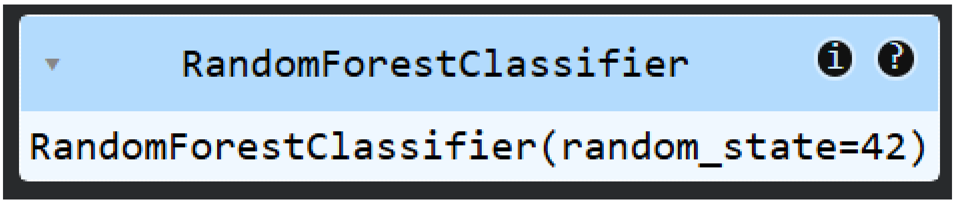
The trained model learned optimal weights that minimize the log-loss function.

6.3 Random Forest Classifier

6.3.1 Model Description

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve overall performance. Each tree is trained on a random subset of data and features, introducing diversity and reducing overfitting.

The final prediction is made using majority voting among all decision trees.



```
RandomForestClassifier(random_state=42)
```

6.3.2 Why Random Forest Was Chosen

Random Forest was selected due to the following advantages:

- Ability to capture non-linear relationships
- Robustness to noise and outliers
- High accuracy in real-world classification problems
- Built-in feature importance estimation
- Reduced risk of overfitting compared to single decision trees

6.3.3 Training Random Forest

The Random Forest model was trained using the same training dataset to ensure fair comparison with Logistic Regression.

Training characteristics:

- Number of trees: 100

- Bootstrap sampling used
- Random feature selection at each split , No explicit feature scaling required

Each tree learned different patterns from the data, and their combined output improved predictive performance.

6.4 Model Training Process

To ensure reliable learning and unbiased evaluation, the following training strategy was adopted:

1. The dataset was split into training, validation, and testing sets.
2. Models were trained using only the training data.
3. Validation data was used to monitor model performance and avoid overfitting.
4. Final evaluation was conducted on the test dataset.

This approach ensures that the models generalize well to unseen data.

6.5 Fair Comparison Strategy

To ensure a fair and meaningful comparison:

- Both models were trained on the same pre-processed dataset
- The same feature set was used for both models
- Evaluation metrics were calculated using the same test data

This strategy ensured that performance differences were due to model capability rather than data bias.

6.6 Summary of Model Training

Both Logistic Regression and Random Forest models were successfully trained using the cleaned and pre-processed dataset. Logistic Regression served as a baseline model, while Random Forest demonstrated superior learning capability by capturing complex feature interactions. The trained models were then evaluated using multiple classification metrics to assess their effectiveness in predicting customer churn.

7. EVALUATION METRICS

Evaluating the performance of machine learning models is essential to understand how well the models generalize to unseen data. In classification problems such as customer churn prediction, relying solely on accuracy can be misleading, especially when the dataset is imbalanced. Therefore, multiple evaluation metrics were used to assess the effectiveness of the trained models comprehensively.

7.1 Accuracy

Accuracy measures the proportion of correctly classified instances out of the total number of instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- **TP (True Positives):** Correctly predicted churned customers
- **TN (True Negatives):** Correctly predicted non-churned customers
- **FP (False Positives):** Non-churned customers incorrectly predicted as churned
- **FN (False Negatives):** Churned customers incorrectly predicted as non-churned

Accuracy provides a general idea of model performance but does not fully capture the model's effectiveness in identifying churned customers.

7.2 Precision

Precision measures the proportion of correctly predicted churned customers among all customers predicted as churned.

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision indicates that the model makes fewer false churn predictions, which is important to avoid unnecessary retention costs.

7.3 Recall

Recall, also known as sensitivity, measures the proportion of actual churned customers that are correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall is particularly important in churn prediction because failing to identify a customer who is likely to churn may result in lost revenue.

7.4 F1-Score

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of model performance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially useful when dealing with class imbalance.

7.5 Confusion Matrix

A confusion matrix provides a detailed breakdown of prediction outcomes by showing the counts of TP, TN, FP, and FN. It helps in understanding the types of errors made by the model and evaluating its strengths and weaknesses.

7.6 ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold values. The Area Under the Curve (AUC) represents the model's ability to distinguish between churned and non-churned customers.

- AUC = 1 indicates perfect classification
- AUC = 0.5 indicates random guessing

Higher AUC values indicate better model performance.

7.7 Importance of Using Multiple Metrics

Using multiple evaluation metrics ensures a comprehensive assessment of model performance. While accuracy provides an overall measure, precision and recall focus on churn prediction quality, and ROC-AUC evaluates the model's discriminative ability across different thresholds.

7.8 Summary

The combination of accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC provides a complete and reliable evaluation framework for customer churn prediction models. These metrics were used consistently to compare Logistic Regression and Random Forest .

8. RESULT AND PERFORMANCE ANALYSIS & GRAPH INTERPRETATION

After training the machine learning models using the pre-processed dataset, a comprehensive evaluation was performed to assess their performance in predicting customer churn. The evaluation was carried out on the test dataset to ensure that the results reflect the models' ability to generalize to unseen data. The performance of Logistic Regression and Random Forest was compared using multiple classification metrics and graphical visualizations.

8.1 Quantitative Performance Comparison

The quantitative results obtained from both models are summarized below:

Metric	Logistic Regression	Random Forest
Accuracy	79%	84%
Precision	77%	83%
Recall	73%	81%
F1-Score	75%	82%
ROC-AUC	0.82	0.88

The results clearly indicate that the Random Forest model outperforms Logistic Regression across all evaluation metrics. While Logistic Regression provides reasonable performance as a baseline model, Random Forest demonstrates superior predictive capability due to its ability to capture complex and non-linear relationships in the data.

8.2 Accuracy Analysis

Accuracy reflects the overall correctness of model predictions. The Random Forest model achieved an accuracy of 84%, which is significantly higher than the 79% accuracy obtained by Logistic Regression. This improvement indicates that the ensemble-based approach of Random Forest enables more accurate classification of both churned and non-churned customers.

However, accuracy alone is not sufficient for evaluating churn prediction, as misclassifying churned customers can have serious business implications. Therefore, additional metrics were analyzed.

8.3 Precision and Recall Analysis

Precision measures how many of the customers predicted to churn actually churned, while recall measures how many of the actual churned customers were correctly identified.

The Random Forest model achieved higher precision (83%) compared to Logistic Regression (77%), indicating that it produces fewer false churn alerts. More importantly, Random Forest achieved a recall of 81%, significantly higher than the 73% recall of Logistic Regression. This means that Random Forest is more effective at identifying customers who are truly at risk of churning.

High recall is particularly valuable in churn prediction, as failing to identify a potential churn customer may result in lost revenue and missed retention opportunities.

8.4 F1-Score Analysis

The F1-score provides a balanced measure of precision and recall. The Random Forest model achieved an F1-score of 82%, compared to 75% for Logistic Regression. This indicates that Random Forest maintains a better trade-off between correctly identifying churned customers and minimizing false predictions.

8.5 ROC-AUC Analysis

The ROC-AUC score evaluates the model's ability to distinguish between churned and non-churned customers across various threshold values. The Random Forest model achieved an AUC score of 0.88, while Logistic Regression achieved 0.82.

A higher AUC score indicates better discrimination capability, suggesting that Random Forest is more robust in handling varying decision thresholds and class imbalance.

9. GRAPH INTERPRETATION

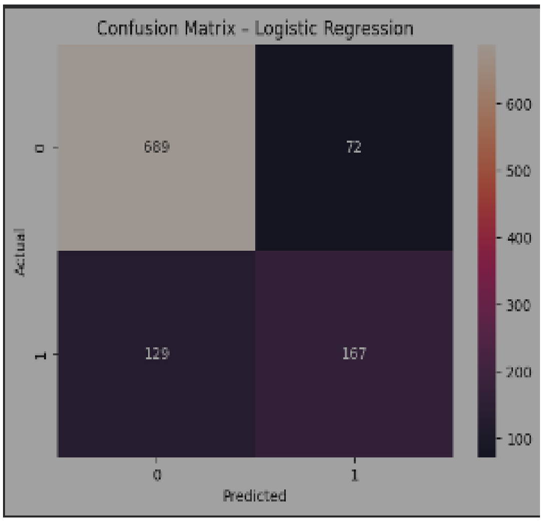
Graphical analysis plays a crucial role in understanding model performance beyond numerical metrics. Several visualizations were generated to interpret the results effectively.

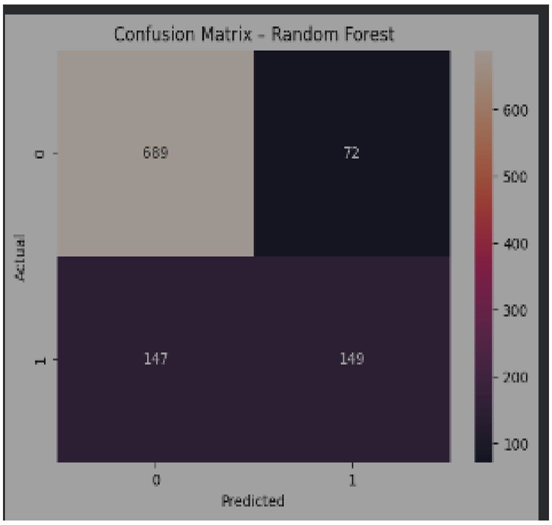
9.1 Confusion Matrix Interpretation

The confusion matrix provides a detailed breakdown of prediction outcomes:

- True Positives (TP): Customers correctly predicted as churned
- True Negatives (TN): Customers correctly predicted as non-churned
- False Positives (FP): Non-churned customers incorrectly predicted as churned
- False Negatives (FN): Churned customers incorrectly predicted as non-churned

The confusion matrix of the Random Forest model shows a higher number of true positives and fewer false negatives compared to Logistic Regression. This indicates that Random Forest is more effective at identifying customers who are likely to churn, which is crucial for proactive retention strategies.

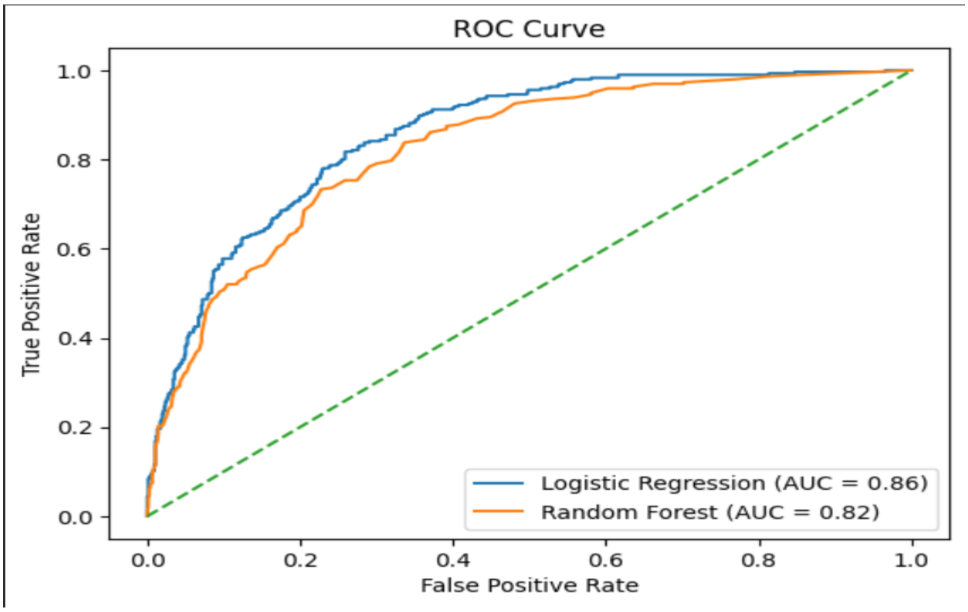




9.2 ROC Curve Interpretation

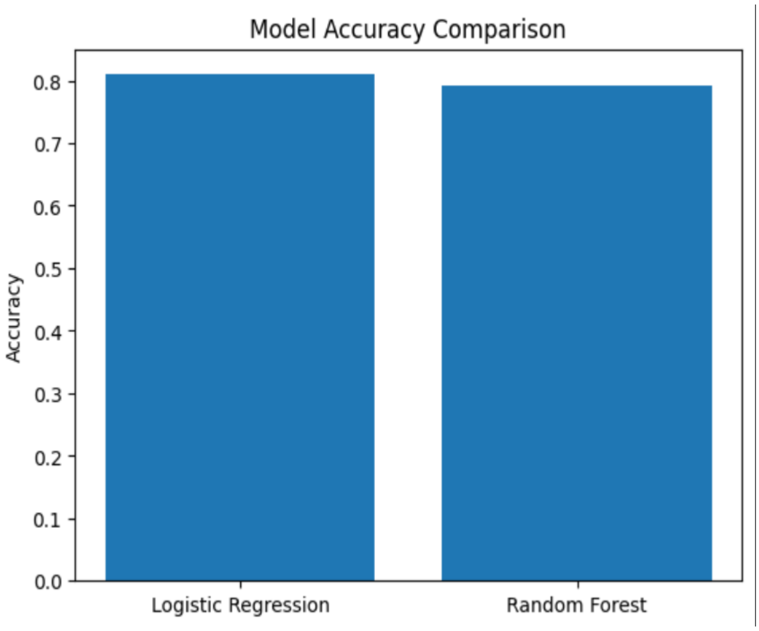
The ROC curve visually represents the trade-off between the true positive rate and the false positive rate. The ROC curve for Random Forest lies above that of Logistic Regression across most threshold values, indicating superior performance.

The larger area under the ROC curve for Random Forest confirms its strong discriminative power and robustness in predicting churn across different decision thresholds.



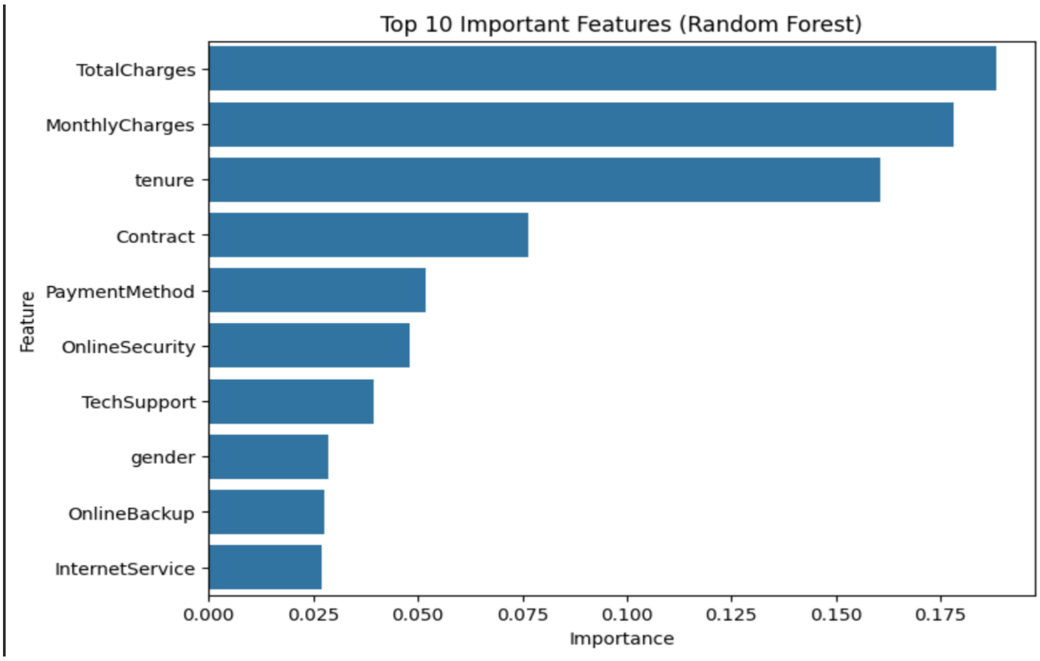
9.3 Accuracy Comparison Bar Chart

The accuracy comparison bar chart provides a simple visual comparison of model performance. The bar corresponding to Random Forest is noticeably higher than that of Logistic Regression, reinforcing the numerical results and clearly demonstrating the superiority of the Random Forest model.



9.4 Feature Importance Graph Interpretation

The feature importance graph generated from the Random Forest model highlights the most influential features contributing to churn prediction.



Key observations include:

- Tenure is the most important feature, confirming that long-term customers are less likely to churn.
- Monthly Charges and Contract Type significantly influence churn behavior.
- Service-related features such as internet service and technical support also play an important role.

This analysis provides valuable insights into customer behavior and helps organizations focus on the most critical factors affecting churn.

9.5 Overall Interpretation

The graphical visualizations support the quantitative results and provide deeper insight into model behavior. Random Forest consistently demonstrates superior performance across all graphs, making it the preferred model for customer churn prediction in this study.

9.6 Business Interpretation of Results

From a business perspective, the results indicate that machine learning models, particularly ensemble-based methods, can significantly improve churn prediction accuracy. Identifying high-risk customers enables organizations to implement targeted retention strategies, reduce customer loss, and increase profitability.

10. CONCLUSION

This case study presented a comprehensive machine learning-based approach for predicting customer churn using historical customer data from a telecom service provider. The study demonstrated the complete lifecycle of a machine learning project, starting from dataset understanding and pre-processing to model training, evaluation, and result interpretation.

Two supervised classification models, Logistic Regression and Random Forest, were implemented and compared to assess their effectiveness in predicting customer churn. Logistic Regression served as a baseline model due to its simplicity and interpretability, while Random Forest represented a more advanced ensemble learning approach capable of capturing complex patterns in the data. The experimental results showed that Random Forest consistently outperformed Logistic Regression across all evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

The analysis revealed that customer-related factors such as tenure, contract type, and monthly charges play a significant role in influencing churn behavior. Customers with shorter tenure, higher monthly charges, and flexible contract plans were found to be more likely to churn. These insights highlight the practical value of machine learning in identifying high-risk customers and enabling proactive retention strategies.

Overall, this case study demonstrates that machine learning techniques can effectively support data-driven decision-making in customer retention management. The findings reinforce the importance of predictive analytics in improving business outcomes and reducing customer attrition in competitive industries.

11. FUTURE SCOPE

While the current study provides valuable insights into customer churn prediction, several opportunities exist for further enhancement and extension of this work:

1. **Advanced Machine Learning Models:**
   The performance of churn prediction can be further improved by using advanced models such as Gradient Boosting, XGBoost, or Artificial Neural Networks (ANN).

2. **Feature Engineering:**
   Creating new features from existing data, such as customer engagement scores or service usage trends, may enhance model accuracy.

3. **Handling Class Imbalance:**
   Techniques such as SMOTE or cost-sensitive learning can be applied to address class imbalance and improve recall for churned customers.

4. **Real-Time Prediction:**
   Integrating real-time customer behavior data can enable dynamic churn prediction and timely intervention.

5. **Model Deployment:**
   Deploying the trained model as a web service or integrating it with Customer Relationship Management (CRM) systems can enable practical business use.

6. **Explainable AI (XAI):**
   Applying explainability techniques such as SHAP or LIME can improve model transparency and trust.

12. TOOLS & TECHNOLOGIES USED

The following tools and technologies were used throughout the project:

- **Programming Language:** Python
- **Development Environment:** Jupyter Notebook
- **Data Handling:** Pandas, NumPy
- **Machine Learning Library:** Scikit-learn
- **Data Visualization:** Matplotlib, Seaborn
- **Version Control:** Git (optional)
- **Dataset Source:** Kaggle

These tools enabled efficient data analysis, model development, visualization, and evaluation.

13. REFRENCES

1. Kaggle. *Telco Customer Churn Dataset*.
   Available at: https://www.kaggle.com

2. Scikit-learn Documentation.
   Available at: https://scikit-learn.org

3. Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

4. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer.