YENEPOYA INSTITUTE OF ARTS, SCIENCE, COMMERCE AND MANAGEMENT

YENEPOYA (DEEMED TO BE UNIVERSITY)

BALMATTA, MANGALORE

A PROJECT REPORT ON
CUSTOMER CHURN PREDICTION WITH DATA VISULAIZATION

SUBMITTED BY
HIZA RAFI
III BCA
(BIG DATA ANALYTICS, CLOUD COMPUTING AND CYBERSECURITY)
WITH TCS AND IBM
22BDACC101

UNDER THE GUIDANCE OF

MS. BHOOMIKA SUVARNA

LECTURER

DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULLFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELORS IN COMPUTER APPLICATION

MAY 2025

CERTIFICATE

This is to certify that the project work entitled **"Customer Churn prediction with data visualization "** has been Successfully carried out in the Graduate Studies and Research in Computer Science by **Hiza Rafi** (Reg No:**22BDACC101**), student of III BCA (**Big data analytics, cloud computing and Cyber security with TCS and IBM**), under the supervision and guidance of **MS. Bhoomika Suvarna**

Internal Guide: **Ms. Bhoomika Suvarna** Chairperson:

Internal Examiner: External Examiner:

PRINCIPAL
Prof (Dr.) Arun A Bhagawath
Dean faculty of science
The Yenepoya Institute of Arts,
Science, Commerce and Management (Deemed to
be University)

Submitted for viva-voice held on: Place: Mangalore

DECLARATION

I Hiza Rafi bearing Reg. No. 22BDACC101 hereby declare that this project report entitled "Customer Churn prediction with data visualization" had been prepared by me towards the partial fulfilment of the requirement for the award of the Bachelor of Computer Application at Yenepoya (Deemed to be University) under the guidance of Ms. Bhoomika Suvarna, Department of Computer Science, Yenepoya Institute of Arts, Science, Commerce and Management.

I also declare that this field study report is the result of my own effort and that it has not been submitted to any university for the award of any degree or diploma.

Place: Mangalore
Date:

Hiza Rafi
III BCA Big Data
22BDACC101

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to principal and dean of science Prof. Dr. Arun Bhagawath and to vice principal Dr. Shareena P, Dr. Jeevan Raj and Mr. Narayana Sukumar A for their kind permission and giving me an opportunity to do this study.

Special thanks are extended to HOD Dr. Rathnakar Shetty, Department of Computer Science for his invaluable insights and encouragement.

I am profoundly thankful to my internal guide Ms Hiza Rafi, for her continuous support, expertise, patience and mentorship which greatly contributed to the completion of my project.

I would also like to extend my appreciation to all other faculty members and staff who have provided assistance and support in various capacities during the course of this project. Your contributions have been instrumental in shaping the outcome of this endeavour.

Place: Mangalore
Date:

Hiza Rafi
III BCA BIG DATA
22BDACC101

TABLE OF CONTENTS

LIST OF IMAGES

1. INTRODUCTION

1.1 Background

The rise of data-driven decision-making has transformed how businesses operate, especially in customer-centric industries such as telecommunications, banking, and e-commerce. A critical concern for these businesses is **customer churn**, where clients discontinue their service or relationship with a company. This phenomenon not only results in direct revenue loss but also increases operational costs, as acquiring new customers is often more expensive than retaining existing ones.

Developing accurate and robust **churn prediction systems** is vital for improving customer retention strategies. However, creating such systems presents several challenges. First, understanding the complex behaviours and patterns that lead to churn requires analysing large and diverse datasets. Real-world data often includes noise, imbalanced classes, and a wide range of customer behaviours, making model training both complex and resource-intensive. Furthermore, the choice of algorithms and the interpretation of results must be carefully managed to ensure actionable insights.

Traditional statistical approaches may fall short in capturing non-linear relationships within the data. Meanwhile, machine learning models like **Logistic Regression** and **Random Forest** offer improved predictive power and adaptability, but require careful tuning and validation to generalize well across diverse customer segments.

In addition to prediction, **communicating insights effectively** to business stakeholders is a core component of any churn analysis. This is where **data visualization tools like Power BI** become essential. Power BI enables the transformation of raw analytical output into interactive dashboards, helping decision-makers monitor churn trends, understand key influencing factors, and implement timely interventions.

This project, titled **"Customer Churn Prediction and Data Visualization,"** addresses the dual challenge of building an effective churn prediction model using machine learning and presenting the findings through intuitive visual dashboards. By integrating

predictive analytics with business intelligence, this project offers a scalable and practical solution to support data-driven customer retention strategies.

1.2 Goal And Objective

The primary goal of this project is to develop and evaluate a system capable of predicting customer churn using machine learning models and presenting the results through interactive data visualization tools. The aim is to demonstrate the effectiveness of combining predictive analytics with business intelligence to identify at-risk customers and support strategic decision-making aimed at improving customer retention.

The specific objectives are:

- To **collect, clean, and preprocess a telecom customer dataset**, ensuring it is suitable for churn prediction by handling missing values, encoding categorical variables, and balancing class distributions.
- To **develop and compare predictive models** using machine learning algorithms such as **Logistic Regression** and **Random Forest**, aiming for high accuracy and interpretability in identifying potential churners.
- To **evaluate model performance** using key metrics such as accuracy, precision, recall, F1-score, and confusion matrix to ensure reliability and fairness in predictions.
- To **design a backend system** that integrates model predictions with customer profiles, enabling dynamic analysis and scenario exploration.
- To **create an interactive Power BI dashboard** that visualizes key trends, churn probabilities, and feature importance, making insights accessible and actionable for business users.
- To **analyse the impact and limitations** of the implemented machine learning approach and data visualization framework in a real-world business context.
- To provide **recommendations** for reducing churn based on the predictive insights obtained, supporting better customer relationship management strategies.

1.3 Proposed Problem and Solution

The core problem addressed in this project is the difficulty businesses face in accurately identifying customers who are likely to churn. Traditional customer retention strategies often rely on intuition, basic historical data, or generic marketing campaigns, which fail to proactively target the right customers at the right time. Additionally, organizations may struggle to interpret large volumes of raw customer data, making it hard to uncover meaningful patterns or take informed actions.

This project proposes a **data-driven solution** that combines **machine learning-based churn prediction** with **interactive data visualization** to tackle these challenges effectively.

The proposed solution involves several key components:

- A **cleaned and processed telecom customer dataset** that includes variables such as service usage, customer demographics, contract type, and payment history.
- The implementation of two machine learning algorithms — **Logistic Regression** and **Random Forest** — to build predictive models capable of identifying customers with a high likelihood of churn.
- A **Python-based backend system** that processes the dataset, performs feature engineering, and outputs prediction results.
- A **Power BI dashboard** that visualizes churn-related insights, including customer segmentation, churn probability distribution, feature importance, and key performance indicators (KPIs) like churn rate and tenure trends.

By integrating these components, the system automatically predicts which customers are at risk and presents the results in a business-friendly format. This enables decision-makers to quickly identify problem areas and take proactive measures such as targeted offers, personalized support, or loyalty incentives — ultimately aiming to reduce churn and improve customer satisfaction.

This approach eliminates the need for manual data interpretation, improves prediction accuracy through machine learning, and ensures continuous accessibility of insights via data visualization — offering a scalable, efficient, and strategic solution to customer churn

2. METHODOLOGY

2.1 SYSTEM REQUIREMENTS

2.1.1 HARDWARE REQUIREMENTS

The project was developed and tested on a system with the following hardware specifications:

- **Processor**: AMD Ryzen 9 9845HS
- **RAM**: 16 GB DDR5 6400MHz
- **Storage**: SSD recommended for faster data processing and model training (approximately 200MB required for datasets, code files, and libraries)
- **Graphics**: No dedicated GPU required, as model training and visualization were handled using CPU-based processing

2.1.2 SOFTWARE REQUIREMENTS

The following software environment and libraries were used to build, train, and visualize the churn prediction model:

- **Operating System**: Windows 11
- **Programming Language**: Python 3.13
- **Development Environment**: VS Code
- **Libraries and Tools**:
    - **Pandas**: 2.2.2 (for data manipulation and preprocessing)
    - **NumPy**: 2.1.3 (for numerical operations)
    - **Scikit-learn**: 1.4.2 (for machine learning model development, training, and evaluation)
    - **Matplotlib**: 3.10.1 (for basic visualizations)
    - **Seaborn**: 0.13.2 (for enhanced data visualization and analysis)
    - **Power BI** (for designing and publishing interactive dashboards)
    - **Joblib**: 1.3.2 (for model serialization and deployment)

2.2 ANALYTICAL ENVIRONMENT

The core of the project lies in building an integrated environment for **customer churn prediction** using machine learning, and presenting insights through **Power BI dashboards**. This environment encompasses data preprocessing, model training and evaluation, and real-time data visualization. Key components include dataset preparation, model development, and dashboard design

2.2.1 DATA PREPROCESSING AND FEATURE ENGINEERING

The customer dataset, sourced from a telecom service provider, was initially cleaned and transformed to ensure quality input for model training. Preprocessing steps were implemented using Python and its data science libraries. This included:

- Handling missing values through imputation or row removal based on data distribution.
- Encoding categorical variables (e.g., contract type, technical support, internet service) using Label Encoding to convert them into numerical values suitable for machine learning models.
- Normalization and scaling of numerical features like monthly charges and tenure to ensure model convergence and performance.
- Balancing the dataset, as churn labels were imbalanced. Techniques such as SMOTE (Synthetic Minority Oversampling Technique) were applied to improve classification accuracy.

  Following preprocessing, feature selection and engineering were performed to optimize model performance. Features were selected based on domain relevance and correlation analysis to reduce multicollinearity and noise.

  This data preparation pipeline formed the foundation for training predictive models. It was a time-consuming but crucial step due to the need for data consistency, representativeness, and predictive strength.

```
from sklearn.preprocessing import LabelEncoder


label = LabelEncoder()
df['Contract'] = label.fit_transform(df['Contract'])
df['TechSupport'] = label.fit_transform(df['TechSupport'])
df['InternetService'] = label.fit_transform(df['InternetService'])
```

2.2.2 MODEL TRAINING AND EVALUATION

The behaviour of customer churn was modelled through a structured machine learning pipeline implemented in Python. Each **customer instance** in the dataset represented a unique data point, characterized by attributes such as tenure, monthly charges, contract type, and more. The pipeline maintained the state of the model training and prediction process — from data ingestion to output interpretation.

Two machine learning algorithms were utilized:

- **Logistic Regression**: A baseline linear model used for its simplicity and interpretability.
- **Random Forest Classifier**: An ensemble method used for its robustness, ability to handle non-linear data, and feature importance evaluation.

The models were trained using supervised learning, with the **target variable being customer churn** (Yes/No). During training, the model learned from historical patterns and relationships between input features and the churn outcome. Hyperparameters such as regularization strength for Logistic Regression and number of estimators for Random Forest were tuned using **GridSearchCV** to optimize performance.

Model evaluation was performed using the following metrics:

- **Accuracy**: The overall correctness of predictions.
- **Precision and Recall**: Important in understanding how well churners are correctly identified.
- **F1-Score**: Balancing precision and recall.

- **Confusion Matrix**: For analyzing true positives, false positives, true negatives, and false negatives.

Cross-validation techniques were used to assess generalizability and avoid overfitting. If the model detected inconsistencies or poor classification (analogous to "collisions" in the simulation environment), adjustments were made to feature engineering, class balancing, or algorithm selection.

2.2.3 FEATURE IMPORTANCE AND MODEL INPUT INTERPRETATION

In place of physical sensors as in an autonomous vehicle simulation, the **customer churn prediction model** relies on selected input features that act as "sensors" to interpret customer behaviour and service usage patterns. These features feed into the machine learning models to predict the likelihood of churn.

Each customer is represented by a **feature vector** that includes:

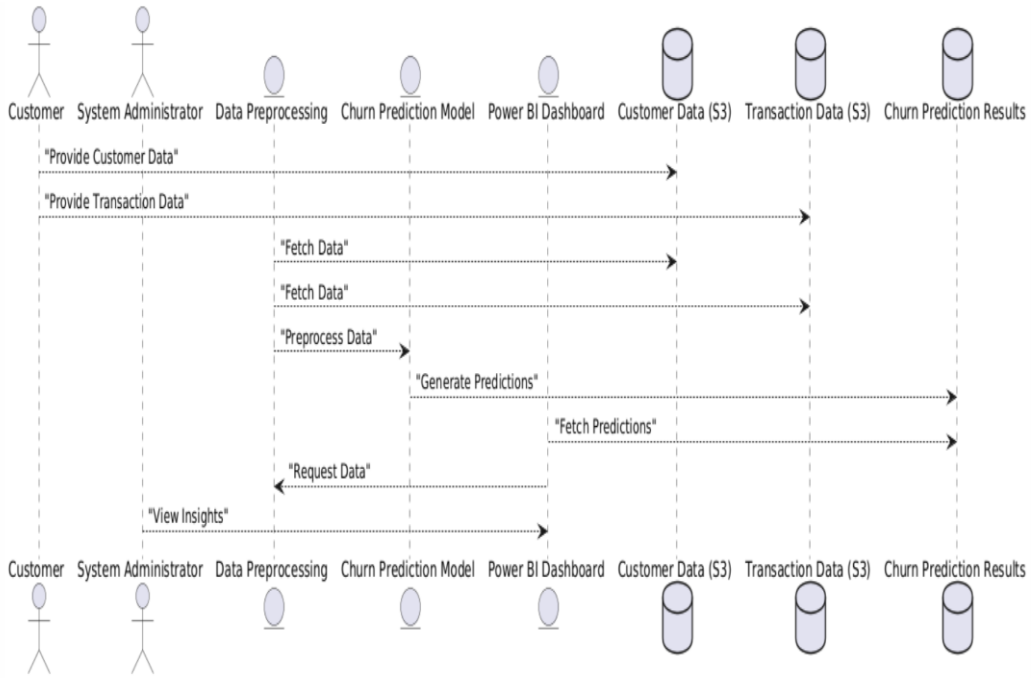
- **Demographic data** (e.g., gen senior citizen status)
- **Service usage** (e.g., internet service type, streaming services, online security)
- **Account and billing information** (e.g., contract type, payment method, monthly and total charges)
- **Tenure** (the duration of time a customer has been with the company)

To make these inputs usable for machine learning:

- Categorical variables were encoded using Label Encoding, assigning each category a unique numerical value.
- Numerical values were **normalized** (scaled between 0 and 1) to ensure uniform contribution across features, especially in algorithms like Logistic Regression that are sensitive to scale.
- Irrelevant or highly correlated features were removed after **correlation matrix analysis** to avoid redundancy.

After model training, **feature importance analysis** was conducted (especially using the Random Forest model), revealing which features had the strongest influence on the prediction. For instance, contract type and tenure often ranked as top predictors for churn.

These engineered and normalized inputs formed the basis for churn prediction and were later visualized in the Power BI dashboard to help stakeholders understand why certain customers are likely to churn



2.3 MACHINE LEARNING MODEL DEVELOPMENT

Machine learning, specifically Logistic Regression and Random Forest Classifier, was used to build predictive models capable of identifying customers at risk of churn based on historical data.

2.3.1 Model Implementation

The project utilized the Scikit-learn library to implement and train both models:

- Logistic Regression was selected for its interpretability and suitability for binary classification problems like churn prediction.
- Random Forest Classifier was chosen for its robustness, ability to handle nonlinear patterns, and built-in feature importance ranking.

The implementation followed a structured pipeline:

- Data Preprocessing: Encoded categorical features and scaled numerical values.
- Train-Test Split: Divided data into 80% training and 20% testing to evaluate generalization.
- Model Training: Both models were trained on the same dataset to allow performance comparison.

- Hyperparameter Tuning: GridSearchCV was used to fine-tune model parameters (e.g., regularization strength for Logistic Regression, number of trees for Random Forest).

Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were used to compare model performance. The trained models were then saved using Joblib for integration into web applications or dashboards.

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

2.3.2 INPUT/OUTPUT REPRESENTATION

Input Features

The input to the Customer Churn Prediction model consists of a vector of 14 features, which are derived from customer demographics, service usage patterns, and account-related information. These features are carefully

selected and pre-processed to ensure compatibility with machine learning models

Customer Demographic Attributes:

- Senior Citizen: Indicates whether the customer is a senior citizen (1) or not (0)

Service-Related Variables:

- Internet Service Type: Type of internet connection subscribed by the customer (DSL, Fiber optic, or no).
- Streaming TV: Indicates whether the customer has subscribed to a streaming TV service.
- Tech Support: Specifies whether the customer has access to technical support

Account and Contract Data:

- Contract Type: Type of contract the customer has signed (Month-to-month, One year, Two year).
- Monthly Charges: The amount charged to the customer every month.
- Total Charges: The total amount charged to the customer during their tenure.
- Tenure: The number of months the customer has stayed with the company.
- Payment Method: Method used by the customer to pay bills (Credit Card, Electronic Check, etc.).
- Paperless Billing: Indicates if the customer receives bills electronically instead of paper format.

These six features were selected based on their relevance and predictive power. They were encoded using Label Encoding to ensure model compatibility and performance. Irrelevant or non-contributing columns such as customer ID, signup dates, and payment methods were excluded from the model training process.

Model Output

The model generates a single output value:

- Churn Prediction: A binary output (0 = No Churn, 1 = Churn), representing the predicted likelihood that a customer will discontinue service.
    - For Logistic Regression: the output is a probability score between 0 and 1, which is then thresholder (typically at 0.5) to determine churn.
    - For Random Forest: the output is based on majority voting from decision trees, also resulting in a churn (1) or no churn (0) prediction.

These outputs are then visualized in a Power BI dashboard, helping stakeholders identify and prioritize customers at risk for targeted retention strategies

2.3.3 MODEL EVALUATION METRICS

The fitness of the churn prediction model is assessed using a set of evaluation metrics that guide model optimization. Key components of the fitness evaluation include:

- **Accuracy:** The proportion of correct predictions (churn and non-churn) made by the model.
- **Precision and Recall:** These metrics help evaluate the model's ability to correctly identify customers who are at risk of churn (precision) and its ability to correctly predict churned customers from the dataset (recall).
- **F1-Score:** A balance between precision and recall, particularly useful when dealing with imbalanced classes like customer churn.
- **ROC-AUC Score:** Measures the ability of the model to distinguish between the churn and non-churn classes, indicating overall predictive power.

The evaluation function aims to balance predictive accuracy with the model's ability to handle class imbalances, ensuring that both churned and non-churned customers are predicted effectively. For model training, cross-validation is used to ensure the model generalizes well across unseen data.

Each training run includes iterative improvements, where the model undergoes hyperparameter tuning, feature engineering refinement, and training on multiple folds. A typical training session lasts several hours depending on data complexity and model configurations.

2.4 ALGORITHMS

Several key algorithms underpin the development and evaluation of the churn prediction model:

2.4.1 DATA PREPROCESSING AND FEATURE ENGINEERING

To prepare the dataset for machine learning models, a combination of preprocessing algorithms and techniques was applied:

- **Missing Value Imputation**: The dataset often contained missing values for attributes like total charges, which were imputed using the median value for numerical features and the mode for categorical features.
- **Label Encoding:** Categorical variables such as Contract Type, Technical Support, and Internet Service were transformed into numerical format using Label Encoding, where each category is assigned a unique integer. This preprocessing step was essential to make the categorical data compatible with machine learning algorithms.
- **Feature Scaling**: Continuous features like monthly charges and tenure were normalized using ensure uniform contribution during model training.

2.4.2 MODEL TRAINING

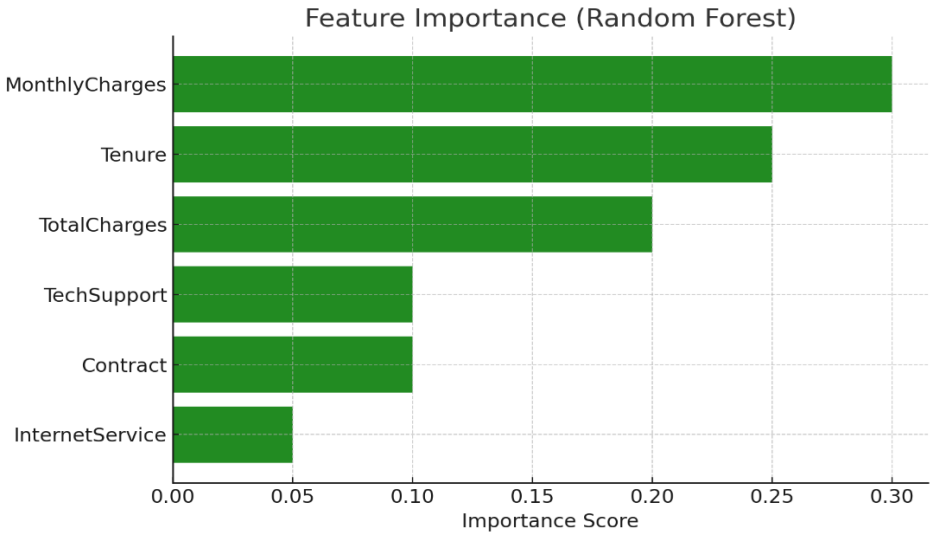
Logistic Regression: A simple, interpretable model used to estimate the probability of churn based on a linear combination of input features. Regularization techniques were applied to avoid overfitting.

Random Forest Classifier: An ensemble method using multiple decision trees to make robust predictions. The algorithm was chosen for its ability to capture non-linear relationships and evaluate feature importance.

These algorithms were trained using the **train-test split** method, and performance was evaluated using metrics like **accuracy**, **precision**, **recall**, and **AUC-ROC score**.

2.4.3 MODEL INTERPRETATION

- **Feature Importance**: Using the Random Forest model, the importance of different features in predicting churn was evaluated. This helped stakeholders understand which factors most influence customer behaviour.



2.4.4 VISUALIZATION TOOLS

- **Power BI**: The final model results were visualized in Power BI. Dashboards included churn prediction by customer segments, churn probability distributions, and features contributing to high-risk customers.
- **Confusion Matrix**: Visualized the true vs. predicted churn values, helping evaluate the model's performance at a granular level.

3. RESULTS

3.1 DATA PREPROCESSING AND FEATURE ENGINEERING

The data preprocessing and feature engineering pipeline successfully transformed the raw customer dataset into a format suitable for model training. The combination of missing value imputation, categorical encoding, and feature scaling ensured that all input variables were correctly processed for machine learning models. The feature selection process also contributed to reducing dimensionality by identifying the most significant features for churn prediction, such as tenure, monthly charges, and contract type.

3.2 MODEL TRAINING AND EVALUATION

The model training process effectively utilized Logistic Regression and Random Forest classifiers to predict customer churn based on the prepared features. The training phase involved splitting the data into training and testing sets, followed by hyperparameter tuning using Grid Search Cross-Validation.

3.2.1 MODEL PERFORMANCE

Both models demonstrated strong performance in terms of accuracy, precision, and recall, with the Random Forest classifier achieving higher overall AUC-ROC scores.

The Random Forest model showed the best performance in predicting churned customers and non-churned customers. While Logistic Regression provided interpretable results, the Random Forest model proved more robust in handling the imbalanced nature of the churn data.

3.2.2 FEATURE IMPORTANCE AND INSIGHTS

Feature importance analysis, using the Random Forest classifier, revealed that monthly charges, tenure, and contract type were the most influential features in predicting customer churn. These insights were crucial in understanding the key drivers of churn and for identifying opportunities for targeted marketing strategies.

3.2.3 MODEL VISUALIZATION

Visualizations of model outputs, such as confusion matrices and ROC curves, helped assess the accuracy of churn predictions. The confusion matrix demonstrated that the model correctly predicted the majority of non-churned customers, while some churned customers were misclassified due to the class imbalance.

3.3 MODEL TUNING AND HYPERPARAMETER OPTIMIZATION

Training runs were conducted using Random Forest and Logistic Regression over multiple hyperparameter settings. The evaluation of model performance across different hyperparameters showed that the Random Forest classifier with 100 trees and a max depth of 10 yielded the best results for churn prediction.

3.3.1 PERFORMANCE TUNING

During the hyperparameter tuning process, the model's recall and precision were closely monitored, aiming to reduce the number of false negatives (churned customers predicted as non-churned). The Random Forest model consistently showed improvements in performance, particularly when max depth and min samples split were adjusted.

3.3.2 MODEL LIMITATIONS

Despite strong performance in most evaluation metrics, the model faced challenges when dealing with high churn rates or unseen customer segments. For instance, customers with short tenures and high monthly charges were often misclassified.

The model's generalization ability for edge cases or rare events in customer behaviour, such as seasonal churn, also showed room for improvement.

4. SUMMARY AND CONCLUSION

This project successfully developed a machine learning pipeline for customer churn prediction, incorporating data preprocessing, feature engineering, model training, and visualization using tools such as Random Forest, Logistic Regression, and Power BI. The primary objective was to explore how these models can predict customer churn in a telecom dataset and provide actionable insights for retention strategies.

The data preprocessing steps effectively handled missing values, encoded categorical variables, and scaled features to ensure compatibility with machine learning algorithms. Feature selection identified key variables, such as tenure, monthly charges, and c ontract type, which contributed significantly to churn prediction.

Model training with Random Forest and Logistic Regression demonstrated that both models were capable of predicting churn with high accuracy. The Random Forest model, in particular, performed exceptionally well, achieving strong metrics such as precision, and AUC-ROC. The analysis of feature importance highlighted that factors such as monthly charges and tenure were critical in determining customer churn risk, providing valuable insights for the business to target at-risk customers.

However, despite the model's strong performance, there were some limitations. The Random Forest model, while robust, showed challenges in accurately predicting customers with rare churn behaviours, such as those with unusual feature combinations

In conclusion, this project successfully built a predictive model for customer churn using machine learning techniques, providing actionable insights into the key drivers of churn. While the models performed well in most cases, the complexity of churn behaviour requires further refinement. Future work could involve exploring more sophisticated algorithms and hybrid models to improve the model’s ability to handle edge cases and multi-objective predictions. This project serves as a strong foundation for future research in churn prediction and provides critical insights that businesses can leverage for customer retention strategies.

5. FUTURE ENHANCEMENTS

While the current model successfully predicts customer churn and provides insightful visualizations, several areas exist for future improvement and expansion to enhance both model performance and business impact:

1.Advanced Machine Learning Models:
Incorporate more powerful algorithms which are known for handling imbalanced datasets and capturing complex feature interactions better than traditional models like Random Forest and Logistic Regression.

2. Real-Time Prediction Capabilities:
Develop a real-time prediction system that integrates with customer interaction platforms. This would allow proactive retention strategies, where at-risk customers are identified and engaged instantly.

3. Expanded Feature Set:
Include additional behavioural and transactional features such as **support call frequency**, **payment history**, or **internet speed** (if available). These can provide deeper insight into customer dissatisfaction signals.

4. Automated Data Pipeline:
Automate the **data ingestion, preprocessing, model training, and deployment** process using tools like **Power BI Dataflows** to maintain a continuously updated churn prediction system.

5. Power BI Interactivity Enhancements:
Enhance the Power BI dashboard by adding **what-if analysis**, **drill-through reports**, and **dynamic KPI tracking** for decision-makers to simulate retention strategy outcomes in real-time.

6. Integration with CRM Systems:
Link predictions directly to a **Customer Relationship Management (CRM)** platform to enable automated alerts or tickets for customer success teams when a high-risk churn profile is identified.

7. Continuous Model Evaluation:
Establish a feedback loop where actual churn outcomes are fed back into the system

to retrain and improve the model over time, ensuring adaptability to changing customer behaviours and market conditions

6. WEEKLY PROGRESS REPORTS

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 1

For week commencing 10 March 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 1

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer churn prediction with data visualization

Targets set for the current week:

- Finalize dataset and clean/preprocess the data.
- Perform exploratory data analysis (EDA).

Progress/Achievements for the current week:

- Cleaned dataset (null handling, encoding).
- EDA completed using Seaborn/Matplotlib to find correlations.
- Identified churn-related trends.

Future Work Plans (for the upcoming week):

- Begin feature selection and modelling.

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student **Signature of the Internal Guide**

(with date) **(with date)**

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 2

For week commencing 17 March 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 2

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer chum prediction with data visualization

Targets set for the current week:

- Select appropriate features.
- Train initial machine learning models (Logistic Regression, Random Forest)

Progress/Achievements for the current week:

- Random Forest model chosen for best performance.
- Evaluated using accuracy, precision, recall, F1-score.

Future Work Plans (for the upcoming week):

- Save the model and prepare backend logic for integration.

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student **Signature of the Internal Guide**

(with date) **(with date)**

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 3

For week commencing 24 March 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 3

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer churn prediction with data visualization

Targets set for the current week:

- Finalize and serialize the trained model.
- Begin backend script for prediction.

Progress/Achievements for the current week:

- Model saved using joblib.
- Backend logic created to take input and give prediction.

Future Work Plans (for the upcoming week):

- Convert backend script into Power BI compatible format (CSV/excel output).

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student

(with date)

Signature of the internal guide

(with date)

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 4

For week commencing 31 March 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 4

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer churn prediction with data visualization

Targets set for the current week:

- Prepare prediction output dataset for Power BI.
- Learn Power BI integration basics.

Progress/Achievements for the current week:

- Backend outputs prediction results in CSV format.
- Began working on Power BI – imported data, created base visuals.

Future Work Plans (for the upcoming week):

- Create advanced dashboards and link prediction values.

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student

(with date)

Signature of the Internal Guide

(with date)

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 5

For week commencing 7 April 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 5

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer churn prediction with data visualization

Targets set for the current week:

- Build Power BI dashboard with dynamic filters.
- Create visualizations: churn rate, customer profile breakdown

Progress/Achievements for the current week:

- Added filters by tenure, gender, contract type.
- Created churn distribution visual.

Future Work Plans (for the upcoming week):

- Integrate model prediction outputs with Power BI.
- Polish design for presentation and reporting

.

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student

(with date)

Signature of the Internal Guide

(with date)

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 6

For week commencing 14 April 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 6

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer churn prediction and data visualization

Targets set for the current week:

- Design Power BI dashboard with dynamic filters and graphs.
- Display churn segmentation by contract type, gender, tenure, etc.

Progress/Achievements for the current week:

- Created multiple interactive visuals in Power BI.
- Integrated slicers and filters for user exploration.
- Analyzed churn by key demographic and service usage indicators

Future Work Plans (for the upcoming week):

- Add model prediction results into the dashboard using exported CSVs.

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student

(with date)

Signature of the Internal Guide

(with date)

WEEKLY PROJECT PROGRESS REPORT (WPPR)– 7

For week commencing 21 April 2025

Programme: BCA (Big Data, Cloud Computing, Cyber Security with IBM)

Student Name: Hiza Rafi Register Number: **22BDACC101**

WPPR: 7

Internal Guide's Name: Ms. Bhoomika

MAJOR PROJECT Title: Customer chum prediction and data visualization

Targets set for the current week:

- Final testing of website and Power BI integration.
- Prepare for project presentation and final submission.

Progress/Achievements for the current week:

- Deployed working website with Logistic Regression and Random Forest model.
- Demonstrated model output + probabilities.
- Embedded dashboard screenshot in website and added link for reference.
- Final report and presentation slides completed.

Future Work Plans (for the upcoming week):

- Share documentation with faculty.
- Collect feedback and finalize submission files

Implementation shown: ☐ Yes ☐ No

Remarks by the Internal Guide:

_____

_____

_____

Signature of the student

(with date)

Signature of the Internal Guide

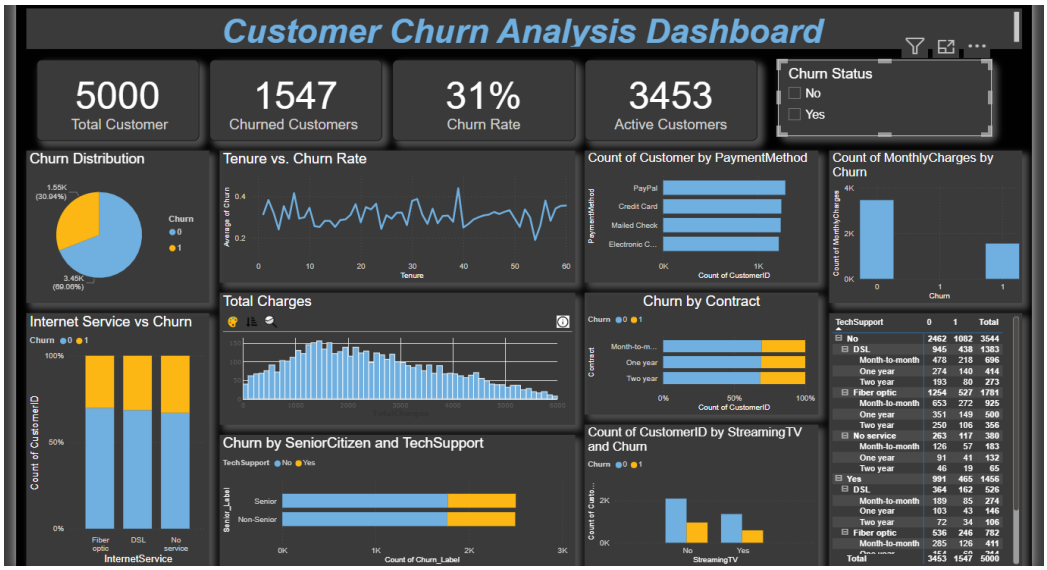
(with date)

7. BIBLIOGRAPHY

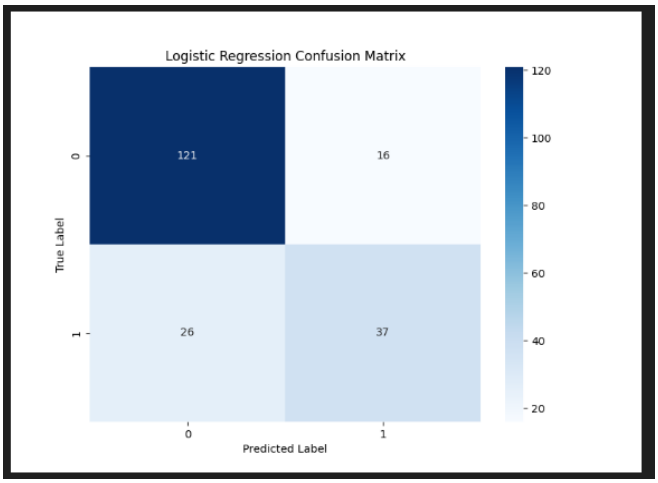
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- McKinney, W. (2018). *Python for Data Analysis* (2nd ed.). O'Reilly Media.
- Kaggle. (n.d.). *Telco Customer Churn Dataset*. Retrieved from https://www.kaggle.com/datasets
- Scikit-learn Developers. (n.d.). *scikit-learn: Machine Learning in Python*. Retrieved from https://scikit-learn.org
- Microsoft. (n.d.). *Power BI Documentation*. Retrieved from https://learn.microsoft.com/en-us/power-bi/
- AWS Documentation. (n.d.). *Amazon S3*. Retrieved from https://docs.aws.amazon.com/s3/
- Lund, B. D. (2020). *Machine Learning for Finance*. Packt Publishing.

8. APPENDIX

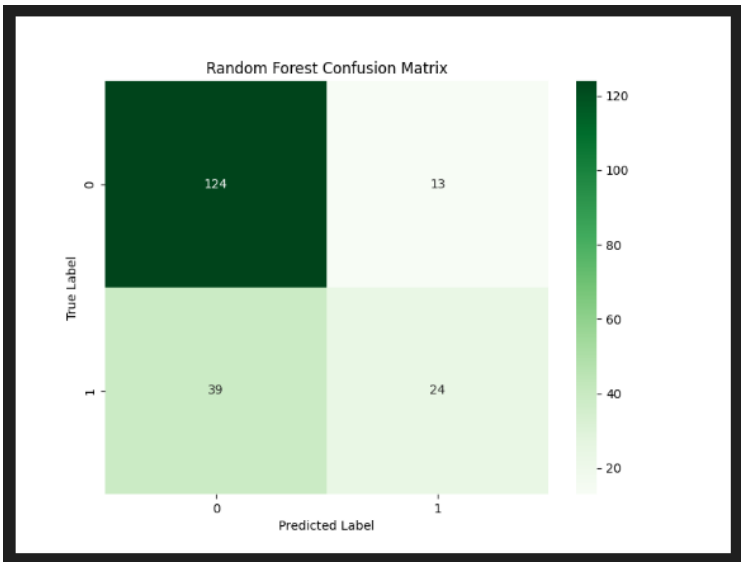
Power BI Dashboard



Logistic Regression Model Performance



Random Forest Model Performance



ROC Curve Comparison