IBM Innovation Centre for Education

PHEMESOFT
Software That Matters
TRUSTED DELIVERY PARTNER for IBM ICE

YENEPOYA
(DEEMED TO BE UNIVERSITY)

YENEPOYA INSTITUTE OF ARTS, SCIENCE AND COMMERCE MANAGEMENT

Customer Churn Prediction with Data Visualization

PROJECT SYNOPSIS

Customer Churn Prediction with Data Visualization

BACHELOR OF COMPUTER APPLICATION

BCA BIG DATA WITH IBM

SUBMITTED BY	GUIDED BY
Mohamed Faheem – 22BDACC147	Sumit K Shukla
Hari Krishnan K Nambiar – 22BDACC095	
Mohammed Shamil T – 22BDACC199	
Muhammad Siyad K – 22BDACC218	

TABLE OF CONTENTS

1. INTRODUCTION

In today's data-driven world, businesses rely heavily on customer retention to maintain profitability. Customer churn—the loss of clients or subscribers—directly impacts a company's revenue and growth. By leveraging the power of data analytics and machine learning, it is possible to identify patterns and factors that contribute to customer attrition and predict which customers are likely to leave.

This project, **Customer Churn Prediction with Data Visualization**, is aimed at building a machine learning-based system that can accurately predict customer churn using historical data. In addition, the project includes the creation of an interactive Power BI dashboard to visualize churn trends, helping business stakeholders make informed decisions.

2. LITERATURE SURVEY

Customer churn prediction has been a popular research topic in both academia and industry for decades due to its direct impact on revenue and customer base. Various machine learning and statistical techniques have been employed to model churn behavior across industries such as telecommunications, banking, e-commerce, and SaaS platforms.

Numerous studies highlight the importance of identifying at-risk customers by analyzing patterns in service usage, customer complaints, payment methods, and contract types. Traditional statistical methods such as logistic regression and decision trees have been widely used, but their predictive power is limited when dealing with complex data patterns. As a result, ensemble learning techniques like **Random Forest**, **Gradient Boosting**, and **XGBoost** have become the preferred choice due to their robustness and higher accuracy.

Research also emphasizes the importance of balancing imbalanced datasets in churn problems. **SMOTE (Synthetic Minority Oversampling Technique)** is one of the most effective methods to handle this, as it synthetically generates new samples for the minority class, resulting in better model training.

In addition to predictive modeling, data visualization tools such as **Power BI** and **Tableau** are increasingly used to provide stakeholders with understandable, interactive insights. Visual dashboards bridge the gap between technical analysis and business decision-making, allowing organizations to act promptly on predictive insights.

This project integrates learnings from these research trends to build a churn prediction system using Random Forest, SMOTE for balancing, and Power BI for visualization — delivering both analytical rigor and user-friendly insight.

3. METHODOLOGY/ PLANNING OF WORK

The project follows a structured and practical approach to identify customer churn patterns and predict churn using data science techniques. Below is the detailed plan:

Phase 1: Data Acquisition

- The dataset used is the **Telco Customer Churn** dataset from Kaggle.
- It includes customer demographics, account information, and service usage data.

Phase 2: Data Preprocessing

- Converted non-numeric columns to numerical format using **Label Encoding**.
- Cleaned the data by handling missing values (e.g., in `TotalCharges`) and removing duplicates.
- Scaled numerical values (`tenure`, `MonthlyCharges`, `TotalCharges`) using **StandardScaler** for model efficiency.

Phase 3: Class Balancing

- Since the dataset was imbalanced (more non-churners than churners), **SMOTE** was applied to create synthetic samples of the minority class, improving model fairness and learning.

Phase 4: Model Development

- Chose **Random Forest Classifier** due to its high performance with tabular data and its ability to handle both numerical and categorical data well.
- The model was trained on 80% of the balanced dataset and tested on 20%.
- Achieved an accuracy of **84%**, with strong recall and precision for both churn and non-churn classes.

Phase 5: Output & Model Storage

- The final dataset with predicted churn labels was saved to a CSV file.
- The trained Random Forest model was saved using **Joblib** for future integration or deployment.

Phase 6: Visualization

- Used **Power BI Desktop** to import the processed dataset and build interactive dashboards.
- Dashboards show churn trends across contract types, payment methods, internet services, and more.
- Visual filters (slicers) and graphs help stakeholders explore data insights interactively.

4. FACILITIES REQUIRED FOR PROPOSED WORK

To complete the project **Customer Churn Prediction with Data Visualization**, both software and hardware infrastructure are necessary for efficient development, training, and visualization. The facilities required can be categorized as follows:

4.1 Hardware Requirements

To handle data processing, machine learning training, and visualization smoothly, the following hardware setup is recommended:

- **Processor:** Intel Core i5 (8th Gen or higher) / AMD Ryzen 5 or equivalent
  (Required for parallel processing and efficient data handling)
- **RAM:** Minimum 8 GB
  (To support loading and transforming large datasets and running ML models in memory)
- **Storage:** 512 GB SSD
  (Fast read/write speed for storing datasets, temporary model files, and tools)
- **Display:** HD resolution (1080p)
  (For Power BI dashboard layout and better visibility of visualizations)

4.2 Software Requirements

The project requires a combination of programming tools for data processing and business intelligence tools for visualization:

- **Operating System:** Windows 10 or above
- **Development Tools:**
  - **Python 3.9 or later** – for scripting and model building
  - **VS Code / Jupyter Notebook** – preferred development environment

- **Python Libraries:**
    - `pandas`, `numpy`, `seaborn`, `matplotlib` – for data manipulation and EDA
    - `scikit-learn` – for model training and evaluation
    - `imbalanced-learn` – for applying SMOTE to handle class imbalance
    - `joblib` – to save the trained model
- **Visualization Tool:**
    - **Power BI Desktop** – to build and present churn trend dashboards
- **Optional (for future enhancements):**
    - **AWS S3** – to store processed data online or integrate for cloud access
    - **Streamlit or Flask** – for deploying the model as a web app (not mandatory)

4.3 Internet and Online Resources

- **Kaggle.com** – For downloading the dataset and accessing similar public datasets

- **Documentation and Tutorials:**

    - scikit-learn, imbalanced-learn, and Power BI documentation
    - YouTube tutorials and GitHub repositories for additional references and guidance.

5. REFERENCES

- Kaggle Dataset – Telco Customer Churn https://www.kaggle.com/blastchar/telco-customer-churn
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.
- imbalanced-learn Documentation: https://imbalanced-learn.org
- Power BI Documentation: https://learn.microsoft.com/en-us/power-bi/