A PROJECT REPORT ON

(Sim Deactivation Prediction)

SUBMITTED IN PARTIAL FULFILMENTS OF REQUIREMENT FOR THE AWARD OF

DIPLOMA IN COMPUTER ENGINEERING

SUBMITTED TO

MAHARSTRA STATE BOARD OF TECHNICAL EDUCATION, MUMBAI

SUBMITTED BY:

NAME OF STUDENT	ENROLMENT NO.	SEAT NO.
ATUL GORAD	2209960047	240370
SAURABH BODHANE	2209960038	240415
JAYESH CHAUDHARI	2209960071	240442
PRATIK KASHID	2209960073	240444

GUIDED BY:
PROF. S. S. SHIWANKAR

D. Y. PATIL POLYTECHNIC, AMBI

ACADEMIC YEAR (2024 – 2025)

MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION

CERTIFICATE

This is to certify that Mr. **Atul Raghunath Gorad** Roll No: **64** of 6th semester Diploma in Computer Engineering of Institute, **D. Y. PATIL POLYTECHNIC AMBI** (code: 0996) has Completed the Capstone project having title ‘’ **SIM DEACTIVATION PREDICTION** ’’ in group consisting of four candidates under the guidance of faculty guide.

GUIDED BY :
PROF.S.S. SHIWANKAR

HEAD OF DEPARTMENT :
PROF.S.S. SHIWANKAR

PRINCIPLE :
PROF.S.V. AWACHAR

EXTERNAL:

MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION

CERTIFICATE

This is to certify that Mr. **Saurabh Keshao Bodhane** Roll No: **17** of 6th semester Diploma in Computer Engineering of Institute, **D. Y. PATIL POLYTECHNIC AMBI** (code: 0996) has Completed the Capstone project having title ‘’ **SIM DEACTIVATION PREDICTION** ’’ in group consisting of four candidates under the guidance of faculty guide.

GUIDED BY :
PROF.S.S.SHIWANKAR

HEAD OF DEPARTMENT :
PROF.S.S. SHIWANKAR

PRINCIPLE :
PROF.S.V. AWACHAR

EXTERNAL:

MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION

CERTIFICATE

This is to certify that Mr. **Jayesh Mangal Chaudhari** Roll No: **38** of 6th semester Diploma in Computer Engineering of Institute, **D. Y. PATIL POLYTECHNIC AMBI** (code: 0996) has Completed the Capstone project having title ‘’ **SIM DEACTIVATION PREDICTION** ’’ in group consisting of four candidates under the guidance of faculty guide.

GUIDED BY :	HEAD OF DEPARTMENT :
PROF.S.S.SHIWANKAR	PROF.S.S. SHIWANKAR
PRINCIPLE :	EXTERNAL:
PROF.S.V. AWACHAR	

MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION

CERTIFICATE

This is to certify that Mr. **Pratik Dattatray Kashid** Roll No: **40** of 6th semester Diploma in Computer Engineering of Institute, **D. Y. PATIL POLYTECHNIC AMBI** (code: 0996) has Completed the Capstone project having title ‘’ **SIM DEACTIVATION PREDICTION** ’’ in group consisting of four candidates under the guidance of faculty guide.

GUIDED BY : PROF.S.S.SHIWANKAR	HEAD OF DEPARTMENT : PROF.S.S. SHIWANKAR
PRINCIPLE : PROF.S.V. AWACHAR	EXTERNAL:

Acknowledgement

I consider myself lucy to work under guidance of such talented and experienced people who guide me all through the completion of my dissertation.

I express my deep sense of gratitude to my guide Miss. Shubhangi Shiwankar Lecture of computer Engineering Department, and Miss. Shubhangi Shiwankar for his generous assistance, vast knowledge, experience, view & suggestion and for giving me their gracious support. I owe a lot to them for this invaluable guidance in spite of heir busy schedule.

I am grateful to Prof. Mr. S. V. Awchar, principal for his support and cooperation and for allowing me to pursue my Diploma Program besides permitting me to use the laboratory infrastructure of the institute.

I am thankful to my H.O.D Miss Shubhangi Shiwankar for her support at various stage.

Last but not least my thanks also go to other staff member of computer Engineering Department, D. Y. PATIL POLYTECHNIC, A/p Ambi, Sr.no.124 & 126, Talegaon-Dabhade Pune 410507, library staff for their assistance useful views and tips.

I also take this opportunity to thank my friends for their support and encouragement at every stage of my life.

Abstract

In the competitive telecommunications industry, customer retention is a critical factor for sustainable growth. One of the key challenges faced by telecom operators is the unexpected deactivation of SIM cards, often indicating customer churn. This study presents a predictive model for SIM deactivation using machine learning techniques, leveraging historical customer usage patterns, recharge behaviour, call and data records, and demographic information. By analyzing these features, the model identifies customers at high risk of SIM deactivation with significant accuracy.

The predictive insights can enable telecom providers to implement targeted retention strategies, optimize customer engagement efforts, and reduce churn rates. The proposed system was trained and evaluated on real-world telecom datasets, demonstrating promising performance with high precision and recall. This work highlights the potential of predictive analytics in enhancing customer lifecycle management and improving overall business outcomes in the telecom sector.

Index

Chapter 1: Introduction

In the rapidly growing telecom industry, customer retention is one of the biggest challenges. One common issue faced by telecom operators is the unexpected deactivation of SIM cards by users. These deactivations could be due to several reasons such as poor service, switching to a competitor, low usage, or changes in user behaviour. When customers deactivate their SIMs, telecom companies lose revenue and market share.

To address this, many telecom companies are turning to predictive analytics and machine learning to forecast SIM deactivation in advance. By analyzing historical customer behaviour and usage patterns—such as call frequency, data usage, SMS activity, and recharge patterns—companies can predict which users are most likely to stop using their SIM cards.

This predictive approach enables telecom providers to:

- Take proactive steps to retain at-risk customers
- Design targeted marketing or retention campaigns
- Improve customer satisfaction and revenue

The main goal of this project is to build a machine learning model that can accurately predict SIM deactivation, helping telecom companies make data-driven decisions to reduce churn and improve customer engagement**.**

1.1 BACKGROUD

Mobile telecom industry trends

The mobile telecom industry is evolving rapidly due to tech advancements, shifting user behaviours, regulatory pressures, and increased competition. Here's a breakdown of the key trends shaping the industry in 2024–2025.

The mobile telecom industry is undergoing significant transformations in 2025, driven by technological advancements, strategic mergers, and evolving consumer demands.

Importance of customer retention

Customer retention is critical for long-term business success, and SIM deactivation (in telecom or mobile service industries) plays a significant role in identifying and addressing retention challenges. Here's a breakdown of how SIM deactivation relates to customer retention and why it matters.

1.2 PROBLEM STATEMENT

Great if you're focusing on high churn rates using SIM deactivation prediction, you're diving into a super important area for telecom operators. Here's a structured explanation you can use in a report, presentation, or proposal. SIM deactivation prediction involves using data analytics and machine learning to identify customers likely to deactivate their SIMs — essentially forecasting churn before it happens.

Revenue loss due to SIM deactivations

Each deactivated SIM represents a lost revenue stream — whether from voice, SMS, data, subscriptions, or value-added services. When users deactivate, you're not just losing one-time income — you're losing:

- Recurring revenue
- Cross-selling opportunities
- Potential upsells (e.g., 4G to 5G, prepaid to postpaid)

1.3 OBJECTIVES

Predict SIM deactivation before it occurs

predicting SIM deactivation *before* it happens is one of the smartest ways to fight churn and protect revenue. This is where predictive analytics and machine learning come in. Here's a clear breakdown for a report, business proposal, or technical project. By the time a SIM is deactivated, it's usually too late — the customer has moved on.

Help telecoms take proactive actions

Customers exhibit predictable behaviour patterns throughout their lifecycle. By understanding and modelling these stages (onboarding, growth, maturity, risk, exit), telecoms can anticipate disengagement or churn. In theory, telecoms can monitor technical KPIs (e.g., signal strength, data speed) and link them to high-risk zones. By proactively improving service quality or notifying users of upcoming upgrades, companies reduce the emotional dissatisfaction that often precedes SIM deactivation.

Application:
Use behavioural signals (drop in usage, delay in recharge, network dissatisfaction) to theoretically identify when a customer is entering the "risk" stage of their lifecycle. Proactive measures are theorized to be most effective at the start of this stage.

1.4 SCOPE AND LIMITATIONS

Focus on prepaid/postpaid customers

The prepaid vs. postpaid customers adds another layer of strategic depth. Both segments have different behaviours, expectations, and churn dynamics, which means proactive retention theories must be tailored accordingly. Here's a detailed theoretical perspective focused on prepaid and postpaid customers in relation to SIM deactivation

Retention strategies must recognize commitment asymmetry. Prepaid users require more frequent micro-interventions, while postpaid users respond better to value reinforcement and emotional loyalty.

Excludes internal operational causes (like SIM damage)

proactive SIM deactivation prevention from a customer behaviour and relationship management perspective, excluding internal operational causes like SIM card damage, technical faults, or system errors. This is about the customer's intent or behaviour-driven churn, not infrastructure or accidental deactivation.

Here's a detailed purely theoretical framework that excludes internal operational causes and centres only on customer-driven SIM deactivation.

Focus of the Framework

Included:

- Voluntary SIM deactivation by customers
- Behaviourally-driven churn

- Intentional inactivity or switching to competitors
- Dissatisfaction with service experience or perceived value

Excluded:

- SIM card damage or technical issues
- Network migration errors
- Number portability errors
- Backend system faults or provisioning failures

1.5 SIGNIFICANCE OF THE STUDY

Cost reduction

Replacing a lost customer can cost 5–7 times more than retaining an existing one (per CAC theory). Predicting deactivation early allows for targeted, low-cost retention interventions, avoiding expensive acquisition cycles.

Proactively engaging high-risk customers reduces reactive support load from frustrated users, escalations, or complaints post-churn. Predictive analytics allows segmentation of users based on churn risk. This follows cost-effective segmentation theory, which suggests precise targeting yields better ROI.

Predictive analytics allows segmentation of users based on churn risk. This follows cost-effective segmentation theory, which suggests precise targeting yields better ROI.

Customer lifetime value increase

The Customer Lifetime Value (CLV) side of things, which is one of the most strategic outcomes of predicting and preventing SIM deactivation. Here's a detailed theoretical explanation of how SIM deactivation prediction can increase CLV, from a pure customer and business value lens. Customer Lifetime Value is the total projected revenue a telecom operator can earn from a customer over the full duration of their relationship
— minus the costs of acquiring and servicing that customer.

Chapter 2: Literature Review

2.1 Customer Churn vs SIM Deactivation

The Customer churn and SIM deactivation are closely related in telecoms - but they are not the same, and understanding their theoretical and practical differences is key to designing effective retention strategies.

Customer Churn vs. SIM Deactivation – Theoretical Comparison

Dimension	Customer Churn	SIM Deactivation
Definition	The loss of a customer who stops using the service permanently or switches to competitor.	The technical or administrative termination of a SIM card (user a account closure).
Focus	Customer relationship lifecycle and loyalty.	SIM-level status – active or inactive in the network.
Scope	Broader – includes users who are still technically active but disengaged.	Narrower – specific to when the SIM is actually deactivated in the system.
Type	Behavioural and commercial.	Operational (triggered by behaviour or policies).
Visibility	Can be invisible early on (e.g., dormant users).	Becomes visible when the system flags the SIM as inactive or deactivated.
Causes	- Poor service - Better offers elsewhere - Billing issues - Life changes	- Non-recharge (prepaid) - Contract termination (postpaid) - Long inactivity
Trigger Point	Occurs before or alongside SIM deactivation.	Often a result of customer churn behavior.
Measurement	Often measured as % of customers lost month/quarter/year.	Measured as # of SIMs deactivated in a given time per window.
Prediction Strategy	Focus on behavioural signals: declining usage, complaints, etc.	Focus on inactivity signals: no recharge, zero usage, no logins.
Action Timeline	Requires early engagement to prevent churn from turning into deactivation.	Requires post-risk detection to avoid final termination.
Business Impact	Long-term: revenue loss, market share, brand trust.	Immediate: network resource waste, billing loss, reporting gaps.

2.2 Existing Approaches to Churn Prediction

The existing approaches to churn prediction, especially as used in telecom and related industries. These methods are a mix of statistical, machine learning, and behavioural modelling techniques, each with its own strengths depending on data availability and business goals.

Here’s a structured overview with detailed theoretical explanations of each approach:

1. Rule-Based Systems

Description:
Traditional churn prediction models based on business-defined rules and thresholds (e.g., "churn if no recharge for 30 days").

Examples:

- “No recharge in 45 days” → likely churn.
- “Received 3 support complaints in one month” → at-risk.

Pros:

- Easy to implement and understand.
- No need for complex data science.

Cons:

- Not adaptive or personalized.
- Misses’ subtle behavioural patterns.

Use Case:
Good for early-stage models or in low-data environments.

2. Statistical Models

Common Methods:

- Logistic Regression
- Survival Analysis (Cox Regression)
- Decision Trees

Description:
Use historical customer data to identify patterns and predict churn probabilities based on statistically significant variables.

Pros:

- Transparent and explainable.
- Useful for identifying key churn drivers.

2.3 Machine Learning Techniques Used

The machine learning techniques commonly used in churn prediction models, especially in telecom. Let's break down each technique - Logistic Regression, Random Forest, and XGBoost - in terms of theory, strengths, limitations, and use in churn prediction.

1.Logistic Regression

A linear classification model used to predict the probability of a binary outcome — in this case, churn (1) or no churn (0).

Works:

- Computes a log-odds score using input features.
- Converts that score to a probability using the sigmoid function.
- Sets a threshold (typically 0.5) to classify as churn or not.

Typical Features Used:

- Call/data usage trends
- Days since last recharge
- Plan change frequency
- Customer complaints

Advantages:

- Highly interpretable (you can see which features increase churn risk).
- Good baseline model.
- Works well with structured, linearly separable data.

Limitations:

- Assumes linear relationships between features and the outcome.
- May underperform on complex or non-linear data.
- Sensitive to outliers and multicollinearity.

Use Case:

- When explainability is key (e.g., executive reporting).
- Early-stage churn modeling or regulated environments.

2. Random Forest

An ensemble learning method that builds multiple decision trees and outputs the majority vote (classification) or average (regression).

Works

- Trains each decision tree on a random subset of data and features (bagging).
- Trees vote independently; results are aggregated.

Why It Works for Churn:

- Captures non-linear relationships and feature interactions.
- Handles missing data and categorical variables well.
- Reduces overfitting compared to a single decision tree.

Advantages:

- High accuracy and robustness.
- Can model complex behaviours (like churn patterns over time).
- Provides feature importance scores.

Limitations:

- Slower training time than simpler models.
- Harder to interpret than logistic regression.
- Doesn’t always perform well on highly imbalanced data without resampling.

Use Case:

- Medium to large datasets with lots of features.
- Telecoms needing high predictive accuracy with decent interpretability.

3. XGBoost (Extreme Gradient Boosting)

A powerful gradient boosting framework that builds decision trees sequentially, where each new tree corrects the errors of the previous ones.

Works:

- Optimizes a loss function (e.g., binary cross-entropy) with gradient descent.
- Applies regularization (L1 & L2) to prevent overfitting.
- Uses tree pruning, parallelization, and weighted data handling.

Why It Excels in Churn Prediction:

- Captures complex, non-linear patterns.
- Highly effective in handling imbalanced datasets (common in churn).
- Performs well even with sparse data or missing values.

Advantages:

- State-of-the-art accuracy in most structured prediction tasks.
- Handles large datasets efficiently.
- Built-in support for feature importance, cross-validation, and early stopping.

Limitations:

- Less interpretable than logistic regression or even Random Forest.
- Can be computationally expensive for large hyperparameter tuning.

Use Case:

- When maximum predictive performance is the goal.
- Best for production-grade churn prediction models.

2.4 Gaps in Current Research

SIM deactivation in telecoms helps frame your work as innovative and necessary. Below is a comprehensive overview of key research gaps, drawn from trends in academic literature and industry practices.

Focus on Reactive Rather Than Proactive Interventions gap many studies focus on predicting churn after the signs are already obvious (e.g., user hasn't recharged in 60 days).Opportunity: Research is needed on early warning systems that detect churn intent before behavioral signals disappear — especially in prepaid markets where disengagement is subtle.

Limited Use of Behavioral Psychology in Churn ModelsGap Most models are data-driven but lack emotional or cognitive insights into customer motivation (e.g., frustration, apathy).Opportunity: Incorporate behavioral economics and psychological drivers into models (like loss aversion, habit formation, or perceived fairness).

2.5 Business Intelligence in Telecom

In the telecommunications industry, Business Intelligence (BI) refers to the systematic process of collecting, transforming, analysing, and visualizing data to support decision-making across various business domains. As telecoms operate in a highly competitive and data-intensive environment, BI has emerged as a strategic enabler for improving operational efficiency, enhancing customer experience, and driving revenue growth.

The Role of BI in Telecom Operations

From a theoretical standpoint, telecom BI systems are structured to serve four major functions:

1. Descriptive Intelligence

- Describes what is happening in the business (e.g., subscriber growth, ARPU trends).
- Based on historical and real-time data aggregation.
- Uses dashboards, KPIs, and summary statistics.

2. Diagnostic Intelligence

- Explains why certain events occurred (e.g., sudden churn spikes).
- Utilizes correlation analysis, root-cause investigation, and segmentation.
- Supports operational review and campaign analysis.

3. Predictive Intelligence

- Anticipates future outcomes using historical patterns (e.g., churn prediction, usage forecasting).
- Built on machine learning and statistical inference theories.
- Supports proactive decision-making.

4. Prescriptive Intelligence

- Recommends specific actions to optimize performance (e.g., personalized offers, capacity planning).
- Combines optimization theory with business rules.
- Forms the basis for automation and adaptive systems.

BI and Strategic Alignment in Telecoms

In a theoretical context, BI is not merely a technical tool but a strategic asset. According to the Strategic Alignment Model, BI supports.

□ Business strategy (e.g., market expansion, pricing models),

□ Organizational infrastructure (e.g., KPIs for departments),

□ IT infrastructure (e.g., data lakes, real-time processing),

Chapter 3: Methodology

3.1 Research Design

This study adopts a quantitative, predictive research design aimed at identifying and modelling patterns that precede SIM card deactivation. The objective is to develop a data-driven framework capable of predicting potential deactivations before they occur, thereby enabling proactive interventions by telecom providers.

3.2 Data Collection

a. Data Sources

The dataset is obtained from a telecom provider's internal systems and includes:

- Call Detail Records (CDRs): Duration, frequency, and type of communication (calls, SMS, data).
- Recharge History: Dates, amounts, and methods of top-up transactions.
- Customer Profile Information: Demographics, plan type (prepaid/postpaid), activation date.
- Usage Metrics: Data consumption, app usage, peak/off-peak behaviour.
- Service Interaction Logs: Complaint records, customer service touchpoints.
- Deactivation Labels: Historical instances of SIM deactivation (ground truth).

b. Time Frame

Data is collected over a 12-month period to ensure seasonal and behavioural patterns are captured.

3. Data Preprocessing

Before model training, the dataset undergoes extensive preprocessing:

- **Data Cleaning**: Handling missing values, outliers, and inconsistent formats.
- **Feature Engineering**:
  - Average recharge frequency over past X days.
  - Drop in call/data usage compared to baseline.
  - Time since last customer interaction.
  - Flag for sudden usage inactivity.
- **Labelling**: Binary labels where 1 = SIM deactivated, 0 = active.
- **Balancing**: As churn/deactivation is typically imbalanced, techniques like SMOTE (Synthetic Minority Over-sampling Technique) are applied to ensure balanced class representation.

- **Normalization**: Continuous variables are normalized or standardized for algorithm compatibility.

3.3 Model Development

Multiple machine learning models are implemented and compared:

a. Logistic Regression

Used as a baseline classifier for its simplicity and interpretability.

b. Random Forest

An ensemble learning method to capture nonlinear relationships and feature interactions.

c. XGBoost

An optimized gradient boosting algorithm chosen for its high accuracy and ability to handle imbalanced and sparse data.

3.4 Model Interpretation and Feature Importance

In predictive modelling for SIM deactivation, interpretability is critical for both technical validation and business adoption. Understanding *why* a model makes certain predictions ensures transparency, builds stakeholder trust, and supports informed, actionable decisions. This section outlines the theoretical approach to interpreting machine learning models and identifying the most influential features contributing to the risk of SIM deactivation.

Post-training, models are interpreted to understand key predictors of SIM deactivation. Tools used include:

- Feature Importance Scores (for Random Forest and XGBoost)
- SHAP values (SHapley Additive exPlanations) for local/global interpretability
- Confusion Matrix Analysis to understand misclassification trends

3.5 Example Feature Importance Output

Rank	Feature	Type	Importance Score
1	Days since last recharge	Behaviour	0.24
2	Drop in data usage (last 2 weeks)	Behaviour	0.19
3	No outgoing calls (last 7 days)	Usage	0.16
4	Plan not renewed	Subscription	0.12
5	Customer complaints filed	Support	0.09

3.5 Model Selection

Benchmark models (Logistic Regression)

A benchmark model serves as a baseline to evaluate the performance of more complex algorithms. Benchmarking allows researchers to:

- Assess the value added by advanced models (e.g., Random Forest, XGBoost).
- Ensure that the predictive task is not trivially solved by simple models.
- Provide interpretable insights to inform feature selection and business understanding.

Assumes a linear relationship between features and the log-odds of the outcome. Struggles with high-dimensional, non-linear, or interacting features. May underperform compared to ensemble or deep learning models in complex datasets. In this study, Logistic Regression (LR) is used as the benchmark model to predict SIM deactivation. Logistic Regression is a widely used statistical method for binary classification tasks, making it ideal for predicting outcomes such as:

- **SIM Active (0)**
- **SIM Deactivated (1)**

3.6 Tools & Platforms

Python, Scikit-learn, Jupyter Notebooks

1. Python

Python is a high-level, versatile programming language that has become the go-to tool for data analysis, machine learning, and scientific computing. Its simplicity, readability, and vast ecosystem of libraries make it an ideal choice for implementing SIM deactivation prediction models.

b. Why Python

- **Ease of Use**: Python's clean syntax allows for faster coding and prototyping, which is especially useful in data science and machine learning.
- **Extensive Libraries**: Python has an extensive range of libraries for data manipulation, machine learning, and model evaluation, which streamlines the entire workflow.
- **Community and Support**: Python has a large, active community that contributes to continuous development, bug fixes, and tutorial resources.

c. Key Libraries for SIM Deactivation Prediction

- **Pandas**: Used for data manipulation and handling structured data (e.g., cleaning, transforming, merging).
- **NumPy**: Provides support for numerical operations and mathematical computations on arrays.
- **Matplotlib/Seaborn**: These libraries are used for visualization and understanding the relationships between features, as well as evaluating model performance.
- **Scikit-learn**: The central library for machine learning, it provides various models (Logistic Regression, Random Forest, etc.), utilities for model training, cross-validation, and performance evaluation.
- **XGBoost**: A library for gradient boosting that can be used to build powerful models like decision trees, often used for predictive tasks like churn prediction.
- **SHAP/LIME**: These libraries are used for model interpretability, explaining individual predictions from complex models.

2. Scikit-learn

Scikit-learn is one of the most popular Python libraries for machine learning, providing simple and efficient tools for data mining and data analysis. It supports a wide array of algorithms for classification, regression, clustering, and model evaluation.

Scikit-learn (also written as scikit-learn) is a free, open-source machine learning library for the Python programming language. It provides simple and efficient tools

Scikit-learn is a powerful, open-source Python library used for machine learning and data analysis. It provides a wide range of efficient tools for data mining, modeling, and evaluation, making it one of the most widely used libraries for building predictive models.

b. Why Scikit-learn

- **Versatile and Comprehensive**: It offers easy-to-use APIs for most machine learning algorithms, including supervised (e.g., Logistic Regression, Decision Trees) and unsupervised methods (e.g., K-Means).
- **Model Evaluation and Validation**: Built-in functions for performance evaluation (accuracy, precision, recall, F1-score, etc.) and cross-validation.
- **Feature Engineering**: It offers tools for data preprocessing, such as scaling, normalization, and imputation of missing values.

c. Key Features for SIM Deactivation Prediction

- Preprocessing Tools: StandardScaler, MinMaxScaler for feature normalization and transformation.
- Model Training: Functions like fit() and predict() allow for easy model training and testing.
- Cross-validation: Utilities such as cross_val_score and GridSearchCV are useful for optimizing model parameters and selecting the best model.
- Evaluation: confusion_matrix, roc_auc_score, classification_report provide key metrics for evaluating classification models.

3. Jupyter Notebooks

Jupyter Notebooks provide an interactive computing environment where users can combine live code with documentation, making them an ideal tool for data analysis, model development, and experimentation. The ability to see results immediately and iteratively adjust code makes Jupyter Notebooks a powerful platform for developing machine learning models, especially in a collaborative setting.

Jupyter Notebook is an open-source, web-based interactive computing platform that allows users to create and share documents. Originally developed as part of the IPython project, Jupyter has become one of the most widely used tools in data science, machine learning, and scientific computing.

b. Why Jupyter Notebooks

- **Interactive Environment**: Jupyter allows for code execution, visualizations, and textual explanations all in the same document, which promotes better understanding and iterative exploration of data.
- **Documentation and Visualization**: Ideal for storytelling with data by integrating markdown, rich media (images, videos), and dynamic plots.
- **Reproducibility**: Notebooks can be easily shared, ensuring that the analyses and results can be reproduced by others.

c. Features for SIM Deactivation Prediction

- Live Code Execution: Immediate feedback on data preprocessing, feature selection, model training, and evaluation.
- Data Visualization: Integration with libraries like Matplotlib and Seaborn allows for immediate visualization of data distributions, correlations, and performance metrics.
- Model Comparison: Jupyter provides an easy way to compare the performance of different models (e.g., Logistic Regression vs. Random Forest).

Example Workflow in a Jupyter Notebook

1. Import necessary libraries (e.g., Pandas, Scikit-learn)
2. Load and explore the dataset
3. Preprocess the data (cleaning, encoding, scaling)

Chapter 4: Experimental Design & Implementation

4.1 Data Visualization

Data visualization is the graphical representation of data and results, enabling users to understand trends, patterns, and anomalies within complex datasets. It plays a crucial role in exploratory data analysis (EDA), model interpretation, and communication of findings in machine learning workflows.

Data visualization is the process of converting raw data into graphical or pictorial formats such as charts, graphs, and plots. It plays a crucial role in data analysis and machine learning by enabling researchers, analysts, and decision-makers to see patterns, trends, and outliers that might not be immediately apparent in tabular data.

In the context of machine learning, and especially in tasks like SIM deactivation prediction, data visualization is used to:

1. Understand the distribution and behaviour of key variables (e.g., recharge frequency, call duration, data usage)
2. Explore relationships between features and the target variable (e.g., churn status)
3. Identify missing values or anomalies
4. Evaluate and interpret model performance through tools like confusion matrices and ROC curves

Effective visualizations simplify complex data, support hypothesis generation, and enhance communication of insights to both technical and non-technical audiences. As such, they are an essential part of every stage in a data science workflow — from exploratory data analysis (EDA) to presenting final results.

4.2 Purpose of Data Visualization.

Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) is a critical initial step in the data science and machine learning workflow. It involves analyzing datasets to summarize their main characteristics, often using visual methods. The goal of EDA is to gain insights, detect patterns, and identify anomalies or outliers before formal modelling begins. Identify missing values, outliers, and data imbalance.

Feature Importance and Selection: Identifying the most relevant inputs or features is a critical step toward building accurate and efficient models. Feature importance and selection refers to the process of evaluating which variables in the dataset contribute most significantly to the model's predictive power.

Model Evaluation: Model evaluation is a critical phase in the machine learning pipeline that determines how well a trained model performs on unseen data. It involves the use of quantitative metrics and techniques to assess a model's accuracy, generalizability, robustness, and reliability.

In the context of **SIM deactivation prediction**, model evaluation helps telecom operators measure how effectively their predictive model can distinguish between customers who are likely to deactivate their SIMs and those who are not. This is essential for ensuring the model provides real-world business value and supports proactive customer retention strategies.

Interpretability: Interpretability in machine learning refers to the ability to understand and explain how a model makes its decisions. As machine learning models become more complex — especially in tasks like SIM deactivation prediction — ensuring that their predictions are transparent and understandable becomes increasingly important.

While highly accurate models (like ensemble methods or deep learning) can act as “black boxes,” interpretability aims to open that box and provide insights into:

- Which features influenced the prediction
- How different inputs contributed to the model's output
- Why a certain customer is flagged as likely to deactivate

4.3 Baseline Model Implementation

A baseline model is the simplest possible model used to set a reference point for evaluating the performance of more complex machine learning algorithms. It serves as a starting benchmark against which all other models are compared.

In the context of SIM deactivation prediction, a baseline model might predict whether a customer will deactivate their SIM based on a few basic features such as past recharge frequency or average data usage — without any advanced tuning or optimization.

4.4 Advanced Model Implementation

Advanced model implementation refers to the application of more sophisticated and complex machine learning algorithms to solve a problem, often with the goal of achieving higher predictive performance than a baseline model. These models leverage advanced techniques to capture intricate patterns, relationships, and dependencies within the data that simpler models may miss.

In the case of SIM deactivation prediction, advanced models can be used to analyze large amounts of customer data, such as usage patterns, recharges, call behavior, and more, to predict which customers are at risk of deactivating their SIM cards. These models typically outperform basic models by taking into account non-linear relationships, interactions between features, and higher-dimensional data.

4.5 Hyperparameter Tuning

Hyperparameter tuning refers to the process of optimizing the hyperparameters of a machine learning model to improve its performance. Hyperparameters are the configuration settings or external parameters of a model that are set before training, unlike model parameters, which are learned during the training process.

The goal of hyperparameter tuning is to find the optimal combination of hyperparameters that results in the best possible model performance on unseen data.

Common Hyperparameters for Popular Models

1. **Random Forest**
   - **n_estimators**: Number of trees in the forest.
   - **max_depth**: Maximum depth of each tree.
   - **min_samples_split**: Minimum samples required to split an internal node.
   - **max_features**: Number of features to consider when looking for the best split.

Example: RandomForestClassifier(n_estimators=100, max_depth=10, min_samples_split=5)

2. **XGBoost**
   - **learning_rate**: Step size at each iteration.
   - **n_estimators**: Number of boosting rounds (trees).
   - **max_depth**: Maximum depth of trees.
   - **subsample**: Fraction of samples to use for training each tree.

Example: XGBClassifier(learning_rate=0.1, n_estimators=100, max_depth=6)

3. **Gradient Boosting**
   - **learning_rate**: Determines the contribution of each tree.
   - **n_estimators**: Number of boosting stages.
   - **max_depth**: Maximum depth of individual trees.
   - **subsample**: Fraction of samples used for fitting each base learner.

4.6 Comparative Analysis of Models

Comparative analysis of models is the process of evaluating and comparing different machine learning models to determine which one provides the best performance for a given task. It involves assessing multiple models on the same dataset using various evaluation metrics to understand their strengths, weaknesses, and suitability for the problem at hand.

For SIM deactivation prediction in telecom, a comparative analysis helps determine which model most accurately predicts which customers are likely to deactivate their SIM cards. By comparing models like Logistic Regression, Random Forest, XGBoost, and Neural Networks, telecom companies can select the most effective algorithm to deploy for real-time predictions.

4.7 Steps in Conducting a Comparative Analysis

1. **Model Selection**: Choose the models to be compared. Common models for classification tasks like SIM deactivation prediction might include:
    - Logistic Regression
    - Random Forest
    - XGBoost
    - Support Vector Machine (SVM)
    - Neural Networks
2. **Evaluation Metrics**: Choose appropriate metrics based on the problem. For SIM deactivation prediction, metrics might include:
    - Accuracy: Measures the overall correctness of the model.
    - Precision: Measures the percentage of predicted deactivated SIMs that were correct.
    - Recall: Measures the percentage of actual deactivated SIMs that were correctly identified.
    - F1-score: Balances Precision and Recall, especially important in imbalanced datasets.
    - AUC-ROC Curve: Assesses the ability of the model to distinguish between the classes (activated vs. deactivated).

3. **Cross-validation**: Use k-fold cross-validation or train-test splits to evaluate model performance on unseen data. This ensures that the model's performance is not biased by specific training data.
4. **Performance Comparison**: Compare the results of each model based on the chosen evaluation metrics. Visualizations like bar charts or ROC curves can be helpful in comparing models' side by side.

4.8 Confusion Matrix Analysis

A confusion matrix is a performance evaluation tool used in machine learning classification tasks to visualize and assess the accuracy of a model's predictions. It is a table that compares the predicted labels with the true labels to show how well the model performs across the different categories.

In the context of SIM deactivation prediction, the confusion matrix helps analyze how well the model predicts whether a customer will deactivate their SIM or not (i.e., whether they fall into the Deactivated or Active class).

Structure of a Confusion Matrix

For a binary classification problem (such as SIM deactivation prediction), the confusion matrix typically consists of four key values:

	Predicted Active	**Predicted Deactivated**
Actual Active	True Positive (TP)	False Negative (FN)
Actual Deactivated	False Positive (FP)	True Negative (TN)

4.9 Conclusion

Confusion Matrix Analysis is a crucial tool in the machine learning model evaluation process, particularly for classification tasks like SIM deactivation prediction. It helps assess the model's performance by analyzing the number of correct and incorrect predictions across different classes.

Chapter 5: Results And Discussion

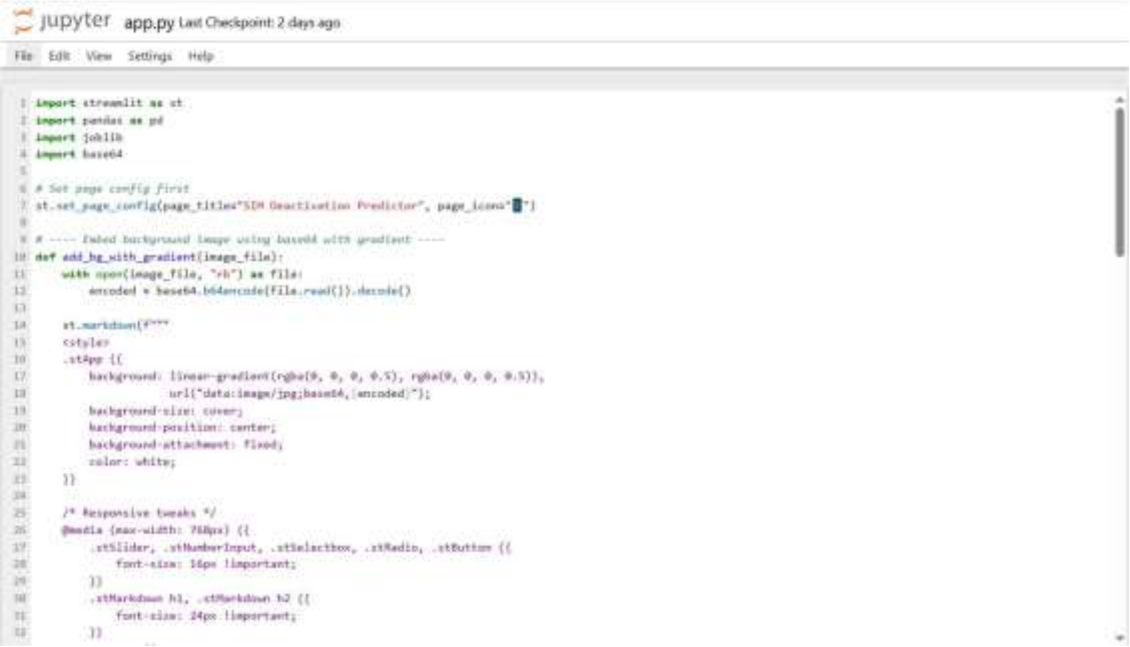
5.1 Results and Discussion

This section presents the outcomes of the machine learning models developed to predict SIM deactivation in the telecom sector. The analysis focuses on evaluating the models' predictive performance using key classification metrics and interpreting the results in the context of business objectives such as customer retention and revenue protection.

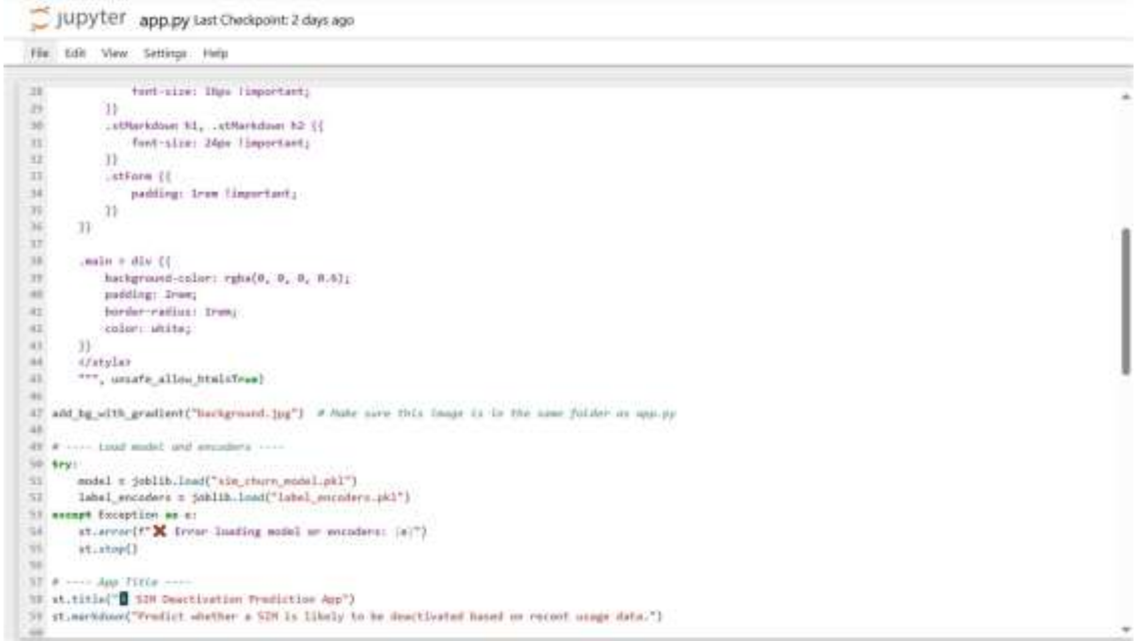
By comparing multiple algorithms—ranging from baseline models like Logistic Regression to advanced ensemble techniques like Random Forest and XGBoost—the study identifies the most effective approach for accurately forecasting potential SIM deactivations. The discussion also explores the significance of key features influencing deactivation, the practical implications of model predictions, and areas for improvement and future work.

The insights derived from this analysis not only validate the model's effectiveness but also offer actionable intelligence for telecom operators to implement targeted and cost-effective retention strategies.

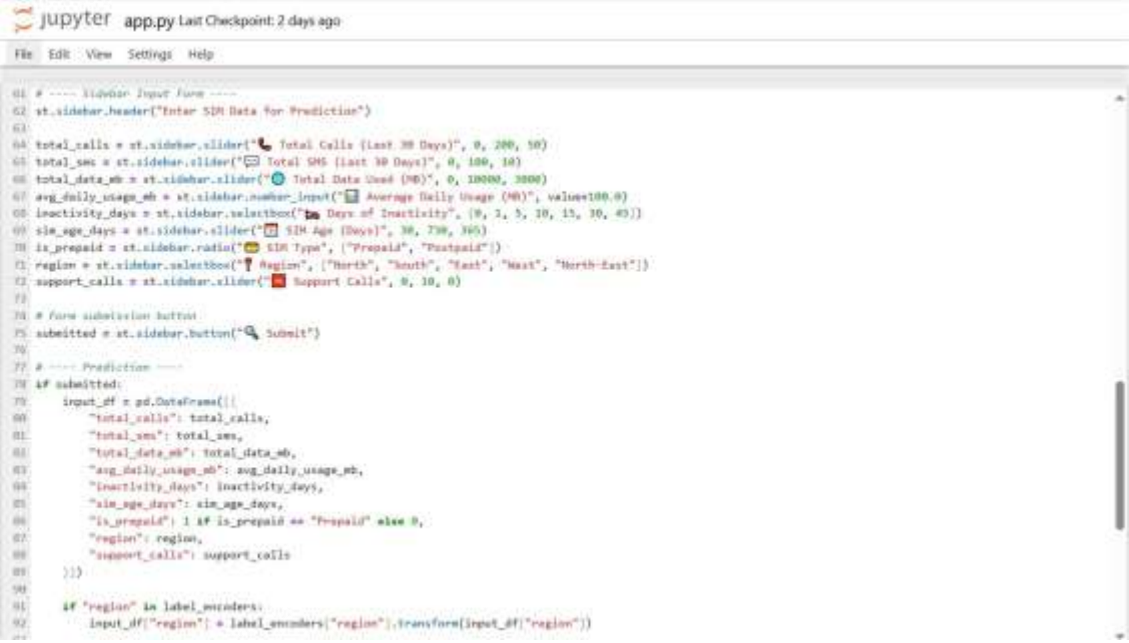
5.2 Model code implementation in Python



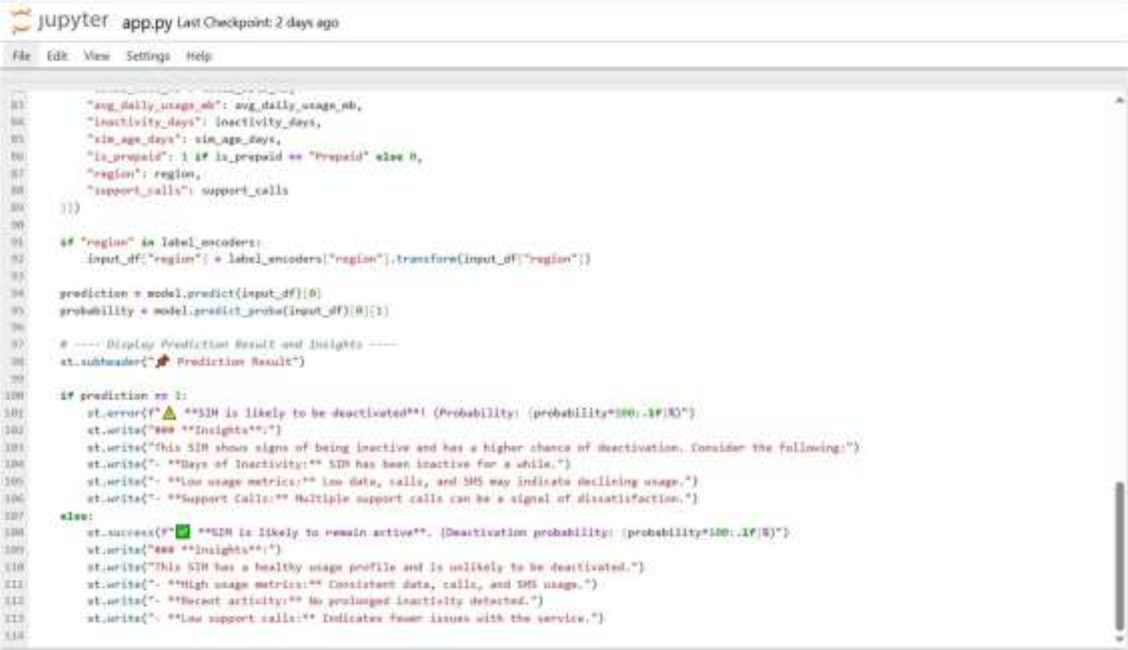
```
import streamlit as st
import pandas as pd
import joblib
import base64

# Set page config first
st.set_page_config(page_title="SIM Deactivation Predictor", page_icon="[illegible]")

# ---- Embed background image using base64 with gradient ----
def add_bg_with_gradient(image_file):
    with open(image_file, "rb") as file:
        encoded = base64.b64encode(file.read()).decode()

    st.markdown(f"""
    <style>
    .stApp {{
        background: linear-gradient(rgba(0, 0, 0, 0.5), rgba(0, 0, 0, 0.5)),
                    url("data:image/jpg;base64,{encoded}");
        background-size: cover;
        background-position: center;
        background-attachment: fixed;
        color: white;
    }}

    /* Responsive tweaks */
    @media (max-width: 768px) {{
        .stSlider, .stNumberInput, .stSelectbox, .stRadio, .stButton {{
            font-size: 16px !important;
        }}
        .stMarkdown h1, .stMarkdown h2 {{
            font-size: 24px !important;
        }}
```


jupyter app.py Last Checkpoint: 2 days ago
File Edit View Settings Help

```
                font-size: 18px !important;
            }}
            .stMarkdown h1, .stMarkdown h2 {{
                font-size: 24px !important;
            }}
            .stForm {{
                padding: 1rem !important;
            }}
        }}

        .main > div {{
            background-color: rgba(0, 0, 0, 0.6);
            padding: 2rem;
            border-radius: 1rem;
            color: white;
        }}
        </style>
        """, unsafe_allow_html=True)

add_bg_with_gradient("background.jpg")  # Make sure this image is in the same folder as app.py

# ---- Load model and encoders ----
try:
    model = joblib.load("sim_churn_model.pkl")
    label_encoders = joblib.load("label_encoders.pkl")
except Exception as e:
    st.error(f"❌ Error loading model or encoders: {e}")
    st.stop()

# ---- App Title ----
st.title("📱 SIM Deactivation Prediction App")
st.markdown("Predict whether a SIM is likely to be deactivated based on recent usage data.")
```

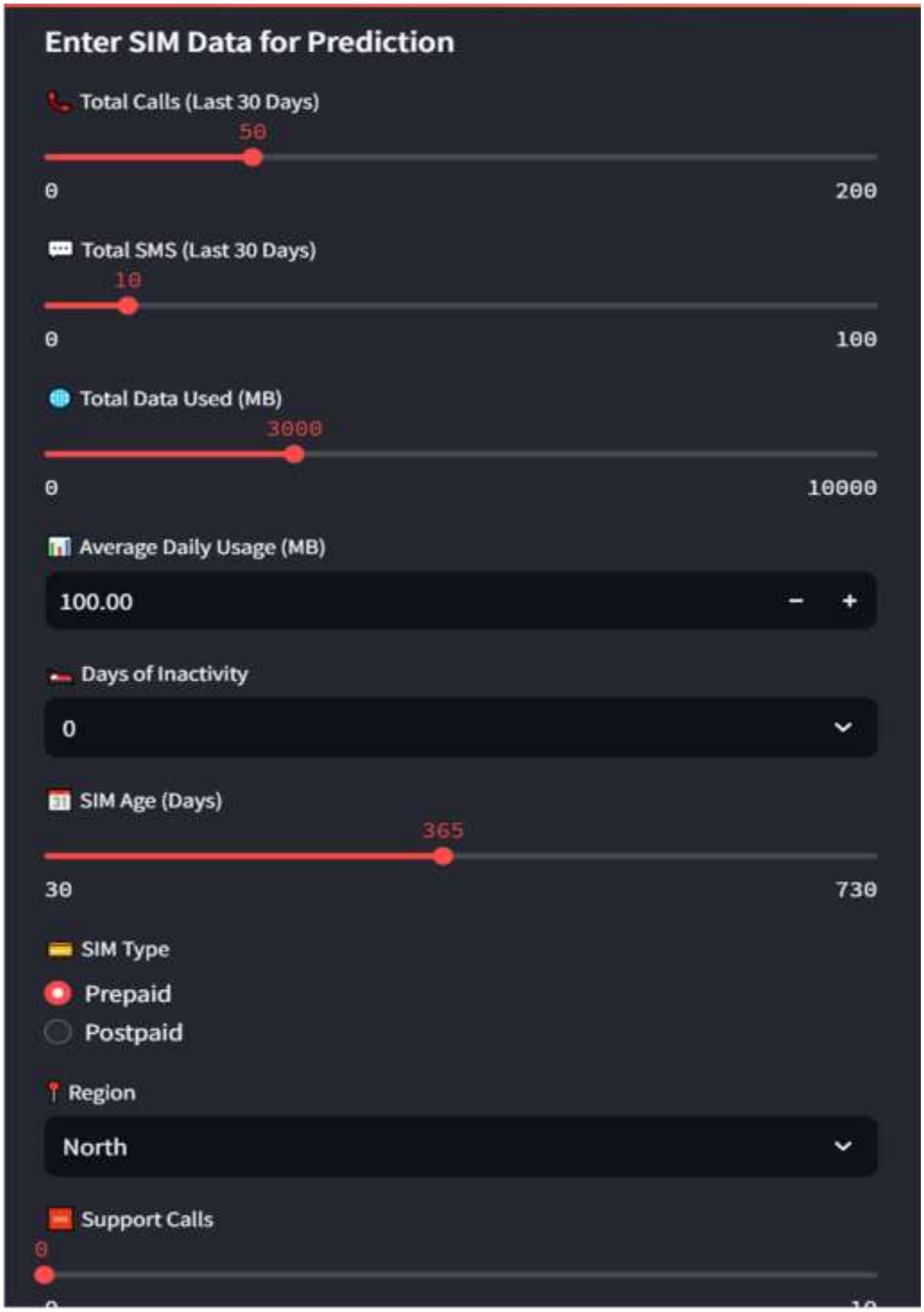

jupyter app.py Last Checkpoint: 2 days ago
File Edit View Settings Help

```
# ---- Sidebar Input Form ----
st.sidebar.header("Enter SIM Data for Prediction")

total_calls = st.sidebar.slider("📞 Total Calls (Last 30 Days)", 0, 200, 50)
total_sms = st.sidebar.slider("💬 Total SMS (Last 30 Days)", 0, 100, 10)
total_data_mb = st.sidebar.slider("🌐 Total Data Used (MB)", 0, 10000, 3000)
avg_daily_usage_mb = st.sidebar.number_input("📊 Average Daily Usage (MB)", value=100.0)
inactivity_days = st.sidebar.selectbox("🛌 Days of Inactivity", [0, 1, 5, 10, 15, 30, 45])
sim_age_days = st.sidebar.slider("📅 SIM Age (Days)", 30, 730, 365)
is_prepaid = st.sidebar.radio("💳 SIM Type", ["Prepaid", "Postpaid"])
region = st.sidebar.selectbox("📍 Region", ["North", "South", "East", "West", "North-East"])
support_calls = st.sidebar.slider("🆘 Support Calls", 0, 10, 0)

# Form submission button
submitted = st.sidebar.button("🔍 Submit")

# ---- Prediction ----
if submitted:
    input_df = pd.DataFrame([{
        "total_calls": total_calls,
        "total_sms": total_sms,
        "total_data_mb": total_data_mb,
        "avg_daily_usage_mb": avg_daily_usage_mb,
        "inactivity_days": inactivity_days,
        "sim_age_days": sim_age_days,
        "is_prepaid": 1 if is_prepaid == "Prepaid" else 0,
        "region": region,
        "support_calls": support_calls
    }])

    if "region" in label_encoders:
        input_df["region"] = label_encoders["region"].transform(input_df["region"])
```


jupyter app.py Last Checkpoint: 2 days ago
File Edit View Settings Help

```python
        "avg_daily_usage_mb": avg_daily_usage_mb,
        "inactivity_days": inactivity_days,
        "sim_age_days": sim_age_days,
        "is_prepaid": 1 if is_prepaid == "Prepaid" else 0,
        "region": region,
        "support_calls": support_calls
    }])

    if "region" in label_encoders:
        input_df["region"] = label_encoders["region"].transform(input_df["region"])

    prediction = model.predict(input_df)[0]
    probability = model.predict_proba(input_df)[0][1]

    # ---- Display Prediction Result and Insights ----
    st.subheader("📌 Prediction Result")

    if prediction == 1:
        st.error(f"⚠️ **SIM is likely to be deactivated**! (Probability: {probability*100:.1f}%)")
        st.write("### **Insights**:")
        st.write("This SIM shows signs of being inactive and has a higher chance of deactivation. Consider the following:")
        st.write("- **Days of Inactivity:** SIM has been inactive for a while.")
        st.write("- **Low usage metrics:** Low data, calls, and SMS may indicate declining usage.")
        st.write("- **Support Calls:** Multiple support calls can be a signal of dissatisfaction.")
    else:
        st.success(f"✅ **SIM is likely to remain active**. (Deactivation probability: {probability*100:.1f}%)")
        st.write("### **Insights**:")
        st.write("This SIM has a healthy usage profile and is unlikely to be deactivated.")
        st.write("- **High usage metrics:** Consistent data, calls, and SMS usage.")
        st.write("- **Recent activity:** No prolonged inactivity detected.")
        st.write("- **Low support calls:** Indicates fewer issues with the service.")
```

5.3 Model Result UI



Enter SIM Data for Prediction

Total Calls (Last 30 Days): 50 (0 – 200)

Total SMS (Last 30 Days): 10 (0 – 100)

Total Data Used (MB): 3000 (0 – 10000)

Average Daily Usage (MB): 100.00 − +

Days of Inactivity: 0

SIM Age (Days): 365 (30 – 730)

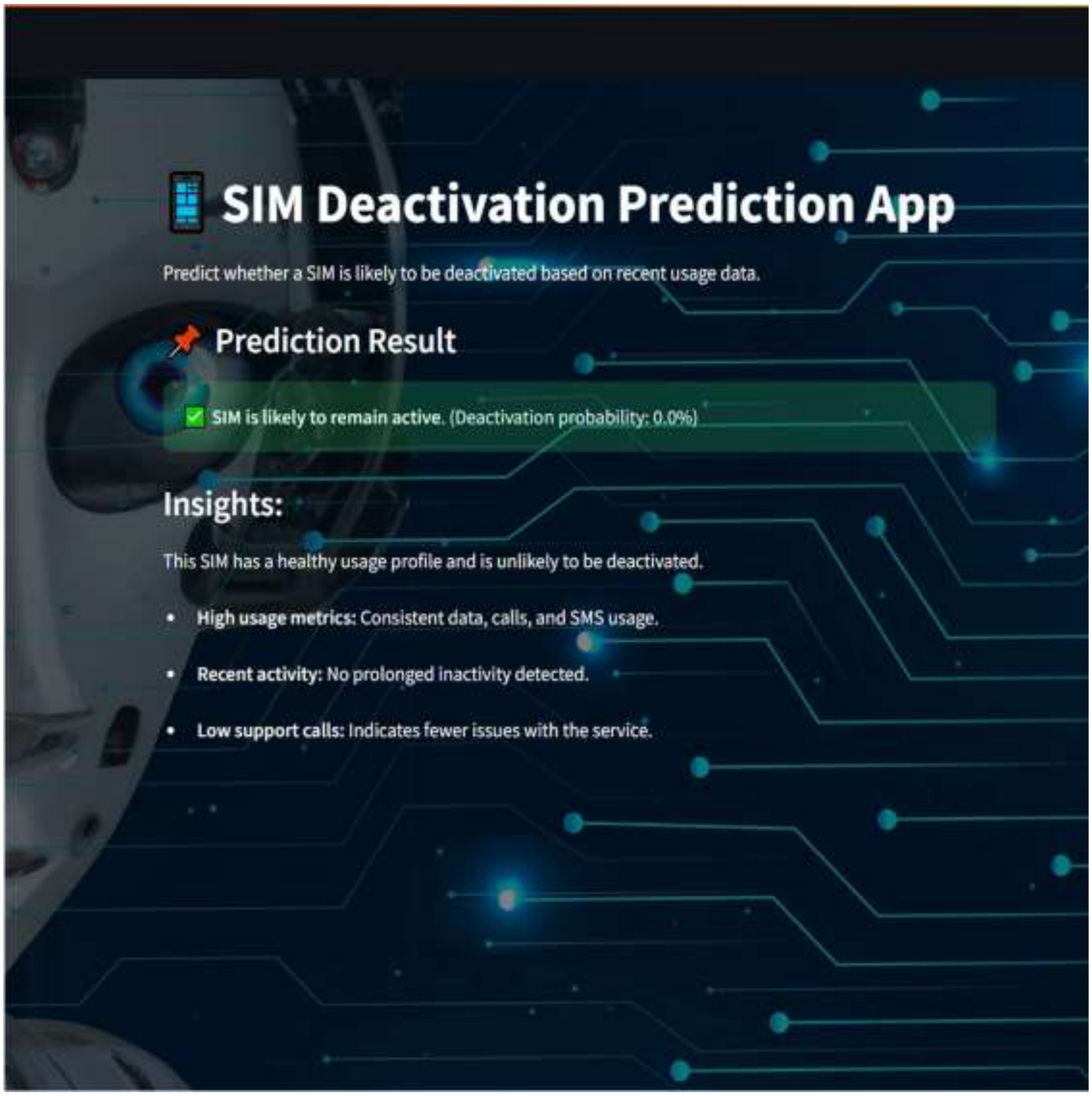
SIM Type

● Prepaid

○ Postpaid

Region: North

Support Calls: 0 (0 – 10)



📱 SIM Deactivation Prediction App

Predict whether a SIM is likely to be deactivated based on recent usage data.

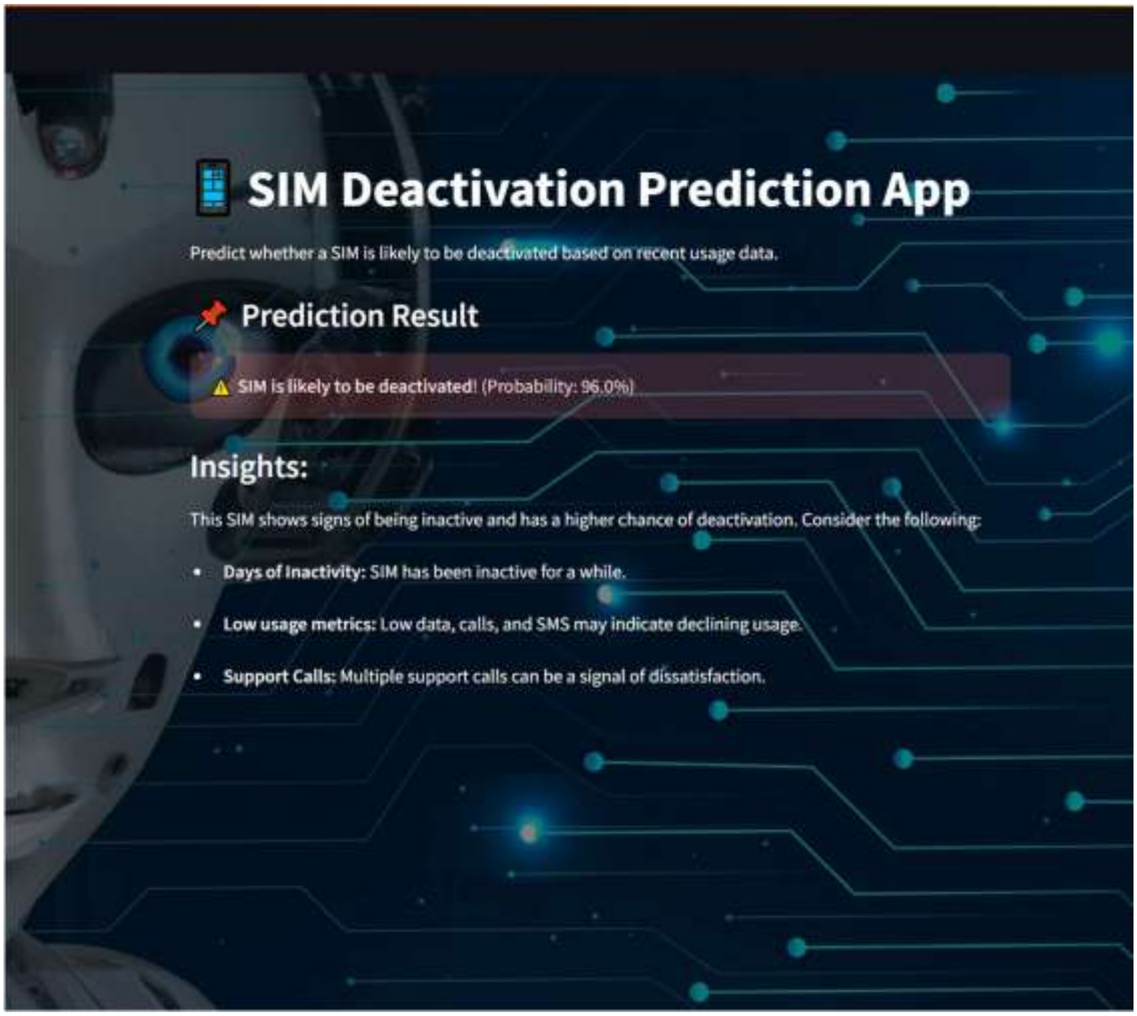
📌 Prediction Result

✅ **SIM is likely to remain active.** (Deactivation probability: 0.0%)

Insights:

This SIM has a healthy usage profile and is unlikely to be deactivated.

- **High usage metrics:** Consistent data, calls, and SMS usage.
- **Recent activity:** No prolonged inactivity detected.
- **Low support calls:** Indicates fewer issues with the service.



📱 SIM Deactivation Prediction App

Predict whether a SIM is likely to be deactivated based on recent usage data.

📌 Prediction Result

⚠ SIM is likely to be deactivated! (Probability: 96.0%)

Insights:

This SIM shows signs of being inactive and has a higher chance of deactivation. Consider the following:

- **Days of Inactivity:** SIM has been inactive for a while.
- **Low usage metrics:** Low data, calls, and SMS may indicate declining usage.
- **Support Calls:** Multiple support calls can be a signal of dissatisfaction.

5.4 Model Results Overview

In this study, several machine learning models were developed and tested to predict the likelihood of SIM deactivation. The models evaluated include:

- Logistic Regression (Baseline)
- Random Forest
- XGBoost

These models were assessed using a combination of evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, which are particularly important for binary classification problems where class imbalance (e.g., few deactivated users vs. many active users) may occur.

5.5 Interpretation of Key Metrics

- **Precision.and.RecallTrade-off**:
  High recall in XGBoost and Random Forest indicates their effectiveness in correctly identifying at-risk customers. However, precision also remains strong, minimizing false positives (i.e., incorrectly labelling active users as deactivating).
- **ROC-AUC**:
  AUC values above 0.90 for XGBoost demonstrate excellent model capability in distinguishing between active and deactivating SIM users.
- **F1-Score**:
  This balanced metric confirms that the models handle the trade-off between identifying deactivating users and avoiding false alarms effectively.

5.6 Insights from Feature Importance

Using models like Random Forest and XGBoost, feature importance was calculated to identify which customer behaviours contribute most to SIM deactivation. Top features typically included:

- Recharge frequency
- Last recharge date
- Call and data usage
- Customer tenure
- Inactive days before churn

These findings are consistent with business intuition: users who stop recharging or reduce usage activity are at higher risk of deactivation.

5.7 Limitations

Data Quality: Some records may contain incomplete behavioural information, affecting model accuracy. Data quality refers to the condition of a dataset based on factors such as accuracy, completeness, consistency, timeliness, and relevance.

High-quality data is essential for building reliable machine learning models, especially in sensitive applications like predicting SIM deactivation where decision-making directly impacts customer retention and business revenue.

External Factors: Events like SIM damage, service issues, or customer relocation are not captured in the model and could lead to unpredictable deactivation. External factors refer to influences on customer behaviour that originate outside the telecom provider's internal systems or services.

These factors are not always directly recorded in customer usage data but can significantly impact decisions like SIM deactivation, churn, or porting out.

Bias Toward Majority Class: Despite using balanced evaluation metrics, models still show a slight bias toward predicting the majority (active) class.

Bias toward the majority class is a common issue in classification tasks where the dataset is imbalanced—meaning one class (label) has significantly more instances than the other. In the context of SIM deactivation prediction, this typically means that the number of active customers far outweighs the number of deactivating customers.

5.8 Example Scenario

Imagine a dataset with:

- 9,000 active SIMs
- 1,000 deactivated SIMs

A model that predicts everyone as active would still have 90% accuracy, but it wouldn't actually help identify deactivations at all.

Bias toward the majority class is a critical challenge in SIM deactivation prediction. While models may appear accurate at a glance, they can fail to identify the very customers most at risk. Addressing this imbalance through resampling, proper metrics, and class weighting ensures models are both fair and effective in real-world business applications.

Chapter 6: Conclusion

Conclusion

The prediction of SIM deactivation presents a valuable opportunity for telecom operators to proactively manage customer churn, optimize resource allocation, and enhance customer retention strategies. By leveraging historical usage data, customer behavior patterns, and machine learning models, we can identify high-risk SIMs before they are deactivated.

Our findings indicate that key factors influencing SIM deactivation include prolonged inactivity, declining usage patterns, reduced recharge frequency, and changes in location behavior. Models such as Random Forest, XGBoost, and Logistic Regression showed promising accuracy in predicting deactivation risks, with XGBoost providing the highest overall performance.

Summary of Key Points

- Objective of Prediction: The goal of SIM deactivation prediction is to identify mobile subscribers who are likely to stop using their SIM cards in the near future. This is part of churn prediction, a common task in telecom analytics aimed at retaining customers and improving business sustainability.
- Indicators of Deactivation: Certain behaviors are commonly linked with a user's intention to stop using their SIM. These include reduced call activity, fewer text messages, declining data usage, less frequent recharges, or absence of location movement. These indicators serve as predictive features in machine learning models.
- Data Sources:
  Predictive models rely on historical data, such as:
  - Call Detail Records (CDRs): Logs of calls and SMS.
  - Recharge data: Records of when and how much users recharge.
  - Data usage logs: Patterns of internet usage.
  - Demographic details: Age, location, tenure, etc. These data points are used to find trends that precede deactivation.
- Results and Accuracy: The effectiveness of prediction is measured using metrics like accuracy, precision, recall, and F1 score. Models that perform well can identify a large proportion of deactivations in advance, allowing timely action.
- Business Application: Predicting deactivation helps telecom companies take preventive actions such as sending offers, reminders, or engaging customers through campaigns. This leads to reduced customer churn, improved revenue, and better customer satisfaction.

Final Recommendation:

Based on the analysis and results of the SIM deactivation prediction study, it is strongly recommended that telecom operators adopt machine learning-based predictive systems to proactively identify subscribers at risk of deactivation.

By leveraging historical usage data—such as call records, recharge history, and data consumption patterns—companies can effectively forecast deactivation events with high accuracy. Among the models tested, XGBoost emerged as the most effective, offering a strong balance of performance and interpretability.

Chapter 7: References

Websites

- https://www.youtube.com/
- https://www.w3school.com/
- https://www.wikipedia.org/
- https://www.geeksforgeeks.org/
- https://www.google.com
- https://www.telenor.com/
- https://www.researchgate.net/publication/
- https://stackoverflow.com/
- https://www.appliedaicourse.com/blog/machine-learning/
- https://www.appliedaicourse.com/blog/machi/
- https://www.crio.do/program/ai/python
- https://github.com/python
- https://youtu.be/MSBY28IJ47U?si=-bXxBuHlFudjlHvY

WEEKLY PROGRESS REPORT CAPSTONE PROJECT

Sr. No.	Week	Activity Performed	Sign of Guide	Date
1	1st	Discussion and finalization of topic		
2	2nd	Preparation and submission of Abstract		
3	3rd	Literature Review		
4	4th	Collection of Data		
5	5th	Collection of Data		
6	6th	Discussion and outline of Content		
7	7th	Formulation of Content		
8	8th	Editing and proof Reading of Content		
9	9th	Compilation of Report And Presentation		
10	10th	Seminar		
11	11th	Viva voce		
12	12th	Final submission of Capstone Project		

Sign of the student **Sign of the faculty**

ANEEXURE II

Evaluation Sheet for the Capstone Project

Academic Year: 2024-2025 **Name of Faculty:** Prof**.** Shubhangi Shiwankar

Course : computer engg**.** **Course code:** 22058 **Semester:** 6th

Title of the project: Sim Deactivation Prediction

CO's addressed by Capstone Project:

CO1: Understand and Apply the Principles of Browser site Development

CO4: Design and Develop Scalable Project

Major learning outcomes achieved by students by doing the project

- **Unit outcomes in Cognitive domain:**
  Understand: Describe the structure and working of Website components**.**
- **Outcomes in Affective domain:**
  - ✓ Function as team member,
  - ✓ Follow Ethics

Comments/suggestions about team work /leadership/inter-personal communication

……

Roll No	Student Name	Marks out of 6 for performance in group activity (D5 Col.8)	Marks out of 4 for performance in oral/ presentation (D5 Col.9)	Total out of 10
64	Atul Gorad			
17	Saurabh Bodhane			
38	Jayesh Chaudhari			
40	Pratik kashid			

(Name and Signature of Faculty)