**Practical Assignment**

| Name | Roshani Patel |
| --- | --- |
| Technology | AI/ML |
| Mentor | Dipam |
| Start Date | 4th August 2025 (Monday) |
| End Date | 8th August 2025 (Friday) |
| Guide | [Interns\_Practical\_Assignment\_Guide](https://docs.google.com/document/d/1pVFKzE-mOdrYuESlg1yh8ZwUvlFWFz6OKaWNHdndWSQ/edit?usp=sharing) |
| Git link | File<Add git link here> |

# **Machine Learning Project: Customer Analytics for a Telecom Provider**

Congratulations on completing your 3-month internship! You have learned a great deal about the fundamentals of Data Science, Machine Learning and Artificial Intelligence. This final assessment is designed to be a capstone project that allows you to apply your knowledge to a realistic, end-to-end problem.

The goal here is not just to achieve the highest accuracy score, but to demonstrate your entire thought process: from understanding the data and business problem to preprocessing, modeling, and interpreting your results. We are most interested in how you approach the problem, justify your decisions, and communicate your findings.

## **1. Domain & Business Context**

**Domain:** Telecommunications

**Problem:** A major telecom company, "ConnectMe," is facing two critical business challenges:

1. **Customer Churn:** They are losing customers to competitors. They need to understand *which* customers are likely to leave (churn). Identifying these customers proactively allows the company to offer them special deals or support to retain them.
2. **Revenue Generation:** To maximize revenue, the company wants to predict the expected monthly spending of new and existing customers. This helps in personalizing marketing campaigns and creating tiered service plans.

Your task is to act as an AI/ML Engineer for a company named "ConnectMe" and build machine learning models to address both of these challenges using a single, unified customer dataset.

## **2. The Dataset:** [**telecom\_customer\_data.csv**](https://drive.google.com/file/d/1010FeaX8R9ps4BxCK-qQu4rtnduzga85/view?usp=drive_link)

You will be provided with a dataset containing anonymized customer information. Here is a description of its columns:

* customerID: Unique identifier for each customer.
* gender: Customer's gender (Male, Female).
* SeniorCitizen: Whether the customer is a senior citizen (1, 0).
* Partner: Whether the customer has a partner (Yes, No).
* Dependents: Whether the customer has dependents (Yes, No).
* tenure: Number of months the customer has stayed with the company.
* PhoneService: Whether the customer has phone service (Yes, No).
* MultipleLines: Whether the customer has multiple lines (Yes, No, No phone service).
* InternetService: Customer's internet service provider (DSL, Fiber optic, No).
* OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies: Additional services (Yes, No, No internet service).
* Contract: The customer's contract term (Month-to-month, One year, Two year).
* PaperlessBilling: Whether the customer has paperless billing (Yes, No).
* PaymentMethod: The customer's payment method (e.g., Electronic check, Mailed check).
* MonthlyCharges: The amount charged to the customer monthly (numeric). **(This will be your regression target)**.
* Churn: Whether the customer churned in the last month (Yes, No). **(This will be your classification target)**.

## **3. Task Definition: A Two-Part Project**

You must complete both of the following tasks using the provided dataset.

### **Part 1: The Regression Task - Predicting Customer Spending**

**Objective:** Build a model to predict a customer's MonthlyCharges. This helps the business forecast revenue and tailor financial plans.

**Target Variable:** MonthlyCharges

**Required Models:** You must implement, train, and evaluate the following three regression models:

1. **Linear Regression:** As a baseline model to understand linear relationships in the data.
2. **Random Forest Regressor:** An ensemble model to capture non-linear interactions between features.
3. **XGBoost Regressor:** A powerful gradient boosting model, often a top performer on tabular data.

**Evaluation Metrics:** For each model, you must calculate and compare:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R-squared (R²) Score

### **Part 2: The Classification Task - Predicting Customer Churn**

**Objective:** Build a model to predict whether a customer will Churn. This is a critical task for customer retention efforts.

**Target Variable:** Churn (You will need to encode this 'Yes'/'No' variable into 1/0).

**Required Models:** You must implement, train, and evaluate the following three classification models:

1. **Logistic Regression:** As a baseline model that provides good interpretability.
2. **Support Vector Classifier (SVC):** A powerful model that works well in high-dimensional spaces. Feel free to experiment with different kernels (e.g., linear, RBF).
3. **Random Forest Classifier:** An ensemble model known for its high accuracy and robustness.

**Evaluation Metrics:** For each model, you must calculate and compare:

* Accuracy
* Precision
* Recall
* F1-Score
* Plot the ROC Curve and report the Area Under the Curve (AUC) score.

## **4. Required Workflow & Deliverables**

You should structure your project in a clear, logical sequence. We recommend using a Jupyter Notebook for your code and analysis.

### **Step 1: Data Exploration and Preprocessing (EDA)**

* Load the dataset.
* Perform exploratory data analysis to understand the distributions and relationships. Use visualizations (histograms, bar charts, correlation heatmaps).
* Handle missing values (if any). Justify your method (e.g., imputation, deletion).
* Encode categorical variables. Explain why you chose a specific encoding method (e.g., One-Hot Encoding vs. Label Encoding).
* Scale numerical features. Justify your choice of scaler (e.g., StandardScaler vs. MinMaxScaler).

### **Step 2: Model Training and Evaluation**

* Split your preprocessed data into training and testing sets.
* For **Part 1 (Regression)**, train all three required models on the training data and evaluate them on the test data using the specified metrics.
* For **Part 2 (Classification)**, train all three required models on the training data and evaluate them on the test data using the specified metrics.

### **Step 3: Hyperparameter Tuning (Optional but Recommended)**

* For at least one complex model from each task (e.g., XGBoost, Random Forest), perform basic hyperparameter tuning (e.g., using GridSearchCV or RandomizedSearchCV) to see if you can improve performance.

### **Step 4: Final Report & Presentation**

Your final deliverable will be a **Jupyter Notebook** and a **short presentation (5-10 slides)** that summarizes your findings. The notebook should be well-commented, and the presentation should cover:

1. **Introduction:** A brief overview of the business problem.
2. **Data Insights:** Key findings from your EDA.
3. **Model Performance:**
   * A clear comparison table for the regression models' performance.
   * A clear comparison table for the classification models' performance.
   * A visualization of the ROC curves for the classification models.
4. **Best Model Selection:**
   * For the regression task, which model would you recommend and why? (Consider performance and other factors like interpretability or speed).
   * For the classification task, which model would you recommend and why? (Justify your choice based on the business problem. Is precision more important, or is recall?).
5. **Conclusion:** A summary of your results and a brief discussion of how "ConnectMe" could use your models.

## **5. Evaluation Criteria**

You will be assessed on the following:

* **Code Quality (20%):** Is your code clean, readable, well-commented, and reproducible?
* **Data Preprocessing & EDA (25%):** Did you make logical choices for handling missing data, encoding, and scaling? Did you explore the data effectively?
* **Model Implementation (25%):** Did you correctly implement and evaluate all six required models?
* **Analysis & Interpretation (30%):** Can you interpret your results? Can you compare the models intelligently and justify your final model choice based on the metrics and the business context? How clearly do you communicate your findings in your report/presentation?