Telco Customer Churn Prediction Project Report

1. Initial Project Description

Objective:
The initial goal of this project was to build a predictive model that can accurately identify whether a customer will churn (i.e., leave the service) based on various customer attributes. Churn prediction is crucial for telecommunication companies as retaining existing customers is generally more cost-effective than acquiring new ones.

Selected Models:
Before analyzing the data, I planned to explore several machine learning models:

- **Logistic Regression**: Chosen for its simplicity and effectiveness in binary classification tasks.
- **Random Forest**: Selected for its robustness and ability to handle complex interactions between features.
- **Gradient Boosting**: Considered for its capability to improve accuracy through sequential learning.

2. Data Description

Dataset Overview:
The dataset consists of 7,043 customer records from a telecommunications company, including 20 features that describe customer demographics, account information, and services used. The target variable is `Churn`, indicating whether the customer has left the service.

Key Features:

- `customerID`: A unique identifier for each customer.
- `gender`: Customer's gender (Male/Female).

- `SeniorCitizen`: Indicates whether the customer is a senior citizen (1 for Yes, 0 for No).
- `Partner`: Indicates whether the customer has a partner.
- `Dependents`: Indicates whether the customer has dependents.
- `tenure`: The number of months the customer has been with the company.
- `PhoneService`: Indicates whether the customer has phone service.
- `MultipleLines`: Indicates whether the customer has multiple lines.
- `InternetService`: Type of internet service (DSL, Fiber optic, No).
- `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`: Indicates whether the customer has subscribed to these additional services.
- `Contract`: The contract term (Month-to-month, One year, Two year).
- `PaperlessBilling`: Indicates whether the customer has paperless billing.
- `PaymentMethod`: The payment method (Electronic check, Mailed check, Bank transfer, Credit card).
- `MonthlyCharges`: The amount charged to the customer monthly.
- `TotalCharges`: The total amount charged to the customer.
- `Churn`: The target variable, indicating whether the customer churned.

Data Types:

- The dataset includes both numerical (`SeniorCitizen`, `tenure`, `MonthlyCharges`, `TotalCharges`) and categorical

(`gender`, `Partner`, `Dependents`, `PhoneService`, etc.) variables.

3. Data Exploration and Cleaning

Exploratory Data Analysis (EDA):

- **Distribution of Churn**: Approximately 27% of the customers in the dataset churned. This indicates an imbalance in the dataset that may need to be addressed during model training.
- **Correlation Analysis**: `tenure` and `MonthlyCharges` are significantly correlated with churn, with customers having shorter tenures and higher monthly charges being more likely to churn.
- **Service Usage**: Customers with no internet service or lower engagement with additional services like `OnlineSecurity` or `TechSupport` tend to churn more often.

Data Cleaning:

- **Handling Missing Values**: The dataset was checked for missing values, and any issues were addressed by filling or dropping rows/columns where necessary.
- **Encoding Categorical Variables**: Categorical variables were converted into numerical formats using techniques like one-hot encoding, particularly for features such as `gender`, `InternetService`, and `Contract`.
- **Scaling**: Numerical features were scaled using StandardScaler to ensure that features like `tenure` and `MonthlyCharges` are on a similar scale for model training.

4. Model Development

Model Selection:

1. **Logistic Regression**: Used as a baseline model due to its simplicity and interpretability in binary classification.
2. **Random Forest**: Chosen for its ability to capture feature interactions and reduce overfitting through ensemble learning.
3. **Gradient Boosting**: Selected for its strength in improving accuracy through sequential training.

Baseline Performance:

- Logistic Regression: Accuracy = 76.97%, AUC-ROC = 0.8157
- Random Forest: Accuracy = 76.05%, AUC-ROC = 0.7961
- Gradient Boosting: Accuracy = 78.39%, AUC-ROC = 0.8296

Cross-Validation and Tuning:

- **Cross-Validation**: Implemented to assess the models' performance across different subsets of the data.
- **Hyperparameter Tuning**: Used GridSearchCV to fine-tune hyperparameters for the Gradient Boosting model, focusing on `learning_rate`, `n_estimators`, and `max_depth`.

5. Model Evaluation

Final Model Performance:

- After tuning, the Gradient Boosting model showed the best performance:
  - Accuracy: 80%
  - AUC-ROC: 0.85

Feature Importance:

- The most important features identified were `Contract`, `tenure`, and `MonthlyCharges`. Customers on month-to-month contracts, with lower tenure and higher monthly charges, were more likely to churn.

Comparison of Models:

- **Logistic Regression**: Provides a solid baseline and is easy to interpret.
- **Random Forest**: Offers robustness and good generalization but performed slightly worse than Gradient Boosting.
- **Gradient Boosting**: Outperformed the other models in both accuracy and AUC-ROC, making it the best choice for this task.

6. Conclusion and Next Steps

Conclusion:

- The Gradient Boosting model was the most effective in predicting customer churn, with a final accuracy of 80% and an AUC-ROC score of 0.85 after tuning.
- This model can be deployed by the company to identify at-risk customers and implement retention strategies, ultimately reducing churn rates.

Next Steps:

- **Model Deployment**: Consider deploying the model into a production environment to make real-time churn predictions.
- **Further Analysis**: Explore additional features or advanced models, such as deep learning, to potentially improve accuracy further.
- **Business Application**: Integrate the model's predictions into customer relationship management (CRM) systems to help in strategizing customer retention efforts.