**Customer Churn Prediction: Documentation and Findings**

**Objective**

The primary goal of this project is to predict customer churn for a telecom company using a synthetic dataset. By accurately predicting which customers are likely to churn, the company can implement strategies to retain these customers and reduce churn rates.

**Exploratory Data Analysis (EDA)**

* **Data Distribution**: I analyzed the distribution of key features such as tenure, monthly charges, and total charges. I observed that customers with shorter tenure and higher monthly charges are more likely to churn.
* **Correlation Analysis**: By creating a correlation matrix, I identified strong relationships between certain features (e.g., Contract type, Monthly charges, and Payment method) and churn. I found that customers with month-to-month contracts and higher charges were at higher risk of churning.

**Data Preprocessing**

* **Handling Missing Values**: Missing data was handled. This step was crucial to ensure that the dataset was clean and ready for model training.
* **Encoding Categorical Variables**: Categorical variables such as Contract type and Payment method were encoded using e.g., one-hot encoding, allowing them to be used in machine learning models.
* **Data Splitting**: The data was split into training and testing sets using a 70-30 split. This allowed me to train the model on one portion of the data and evaluate its performance on unseen data.

**Feature Engineering**

* **Derived Features**: I created new features, such as average\_monthly\_charge, to enhance the predictive power of the models.
* **Feature Selection**: Based on the correlation analysis and preliminary model results, I selected features that were most relevant to predicting churn, such as Contract type, Tenure, and Monthly charges.

**Model Building**

* **Model Choices**: I experimented with several models, including Logistic Regression, Decision Tree, and Random Forest, to determine which provided the best performance.
  + **Logistic Regression**: Chosen for its simplicity and ease of interpretation.
  + **Decision Tree**: Selected for its ability to capture non-linear relationships.
  + **Random Forest**: Used for its robustness and ability to handle feature interactions effectively.
* **Hyperparameter Tuning**: I performed Grid Search to optimize the models' hyperparameters, which helped improve their performance.

**Model Evaluation**

* **Comparison of Models**: Each model was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Random Forest performed the best, with an accuracy of C% and the highest ROC-AUC score.
* **Confusion Matrix Analysis**: Confusion matrices were generated for each model to understand the types of errors made. Random Forest showed the best balance between false positives and false negatives.
* **ROC-AUC Curves**: The ROC-AUC curves confirmed that Random Forest provided the best trade-off between sensitivity and specificity.

**Model Selection**

* **Final Model Choice**: Based on the evaluation metrics and business context, I selected the Random Forest model as the final model. It provided the highest accuracy and the best overall performance.
* **Feature Importance**: Analyzing the feature importance in the Random Forest model revealed that Contract type, Tenure, and Monthly charges were the most influential features in predicting churn.

**Key Findings**

* Customers with month-to-month contracts and higher monthly charges are more likely to churn.
* The Random Forest model provided the best accuracy and overall performance in predicting churn.

**Recommendations**

* The company should focus on retaining customers with month-to-month contracts by offering incentives for long-term contracts.
* Reviewing the pricing structure for high monthly charges could help reduce churn.
* Future improvements could involve exploring more advanced models like Gradient Boosting or incorporating additional data sources.

**Conclusion**

This project successfully developed a model to predict customer churn, with the Random Forest model being the best-performing model. The insights gained from this analysis can help the company implement targeted strategies to reduce churn and retain customers.