Telco Customer Churn Analysis- Technical Report

The task is to prepare a comprehensive report on the Telco Customer Churn prediction project. This report should include:

- **What Happened (Project Scope)**: Detail the project's objective of predicting customer churn, the use of the "Telco Customer Churn" dataset located at "/content/WA_Fn-UseC_-Telco-Customer-Churn.csv", and the analysis steps undertaken. These steps include data cleaning (handling `TotalCharges` type conversion, verifying no nulls/duplicates), exploratory data analysis (correlations, churn distribution, impact of `MonthlyCharges`, `gender`, `tenure`, `Contract`, and other categorical features on churn), feature engineering (creating `MonthlyCharges_Category_Custom`), model training (Logistic Regression, Random Forest, XGBoost), and model evaluation (accuracy, classification reports, confusion matrices).
- **Who Cares?**: Identify key stakeholders within a telecom company who would benefit from this churn analysis, such as management, marketing, customer service, and product development teams.
- **Questions**: State the primary questions the analysis aimed to answer, including identifying key churn drivers, accurately predicting customer churn, and comparing the performance of different predictive models.
- **Goal**: Clearly define the main objective of the project: to build a system that accurately predicts customer churn to enable the implementation of proactive customer retention strategies.
- **What they want to know (Metrics)**: List the key performance indicators used for model evaluation, specifically Accuracy, Precision (for both 'No Churn' and 'Churn' classes), Recall (for both 'No Churn' and 'Churn' classes), F1-score (for both 'No Churn' and 'Churn' classes), Macro F1, and Weighted F1.
- **What success looks like**: Describe the criteria for a successful project outcome, such as achieving acceptable prediction accuracy, deriving actionable insights from the analysis, and developing a deployable churn prediction model.
- **What is the impact**: Explain the potential business benefits of implementing this churn prediction system, including reduced customer attrition, improved customer retention, more targeted marketing campaigns, and enhanced customer satisfaction.
- **Recommendation**: Provide specific recommendations based on the model comparisons and feature importance analysis. Recommend which model (Random Forest or XGBoost) is suitable for deployment in production for predictive tasks and which model (Logistic Regression) is best for explaining churn drivers to stakeholders due to its interpretability. Highlight the most significant churn drivers identified (e.g., Contract type, tenure, Monthly Charges, Online Security, Tech Support, Internet Service Type, Senior Citizen, Paperless Billing) and suggest actionable strategies for customer retention based on these insights (e.g., focusing on customers with month-to-month contracts, senior citizens, and those using paperless billing).

Key Stakeholders and Their Interests in Churn Analysis:

1. **Executive Management (CEO, COO, General Managers):**
    - **Interest:** Overall business performance, revenue stability, customer base growth, and strategic planning. Churn directly impacts these metrics.
    - **How they use it:** To set company-wide strategies for customer retention, assess the financial impact of churn, evaluate the effectiveness of retention initiatives, and make high-level decisions regarding investment in customer experience or new service offerings.
2. **Marketing Department:**
    - **Interest:** Customer acquisition costs, campaign effectiveness, customer segmentation, and targeted retention campaigns.
    - **How they use it:** To identify at-risk customer segments for targeted retention campaigns, personalize offers to prevent churn, refine new customer acquisition strategies by understanding churn drivers, and improve customer messaging.
3. **Customer Service Department:**
    - **Interest:** Customer satisfaction, service quality, complaint resolution, and reducing inbound churn-related calls.
    - **How they use it:** To proactively identify customers likely to churn based on their interactions, prioritize support efforts for high-value at-risk customers, train agents on common churn reasons, and improve service processes based on churn feedback.
4. **Product Development / Product Management Teams:**
    - **Interest:** Product usage, feature adoption, identifying product gaps, and developing services that enhance customer loyalty.
    - **How they use it:** To understand which product features contribute to churn or retention, identify unmet customer needs, prioritize new feature development or product enhancements, and ensure products align with customer expectations.
5. **Sales Department:**
    - **Interest:** Customer retention, upselling/cross-selling opportunities, and understanding customer lifetime value.
    - **How they use it:** To focus on retaining existing customers, identify opportunities to offer bundled services or upgrades that might increase loyalty, and understand customer segments that are more likely to stay.
6. **Finance Department:**
    - **Interest:** Revenue forecasting, budget allocation for retention efforts, and understanding the financial implications of customer attrition.
    - **How they use it:** To forecast future revenue more accurately by accounting for anticipated churn, justify spending on retention programs, and evaluate the ROI of customer loyalty initiatives.

Primary Questions Addressed by the Analysis

The customer churn prediction analysis aimed to answer the following core questions:

1. **What are the key drivers of customer churn?** This involves identifying which features (e.g., contract type, tenure, monthly charges, internet service, security, support) have the most significant impact on a customer's decision to leave the telecom company.
2. **Can we accurately predict which customers are likely to churn?** This addresses the effectiveness of machine learning models in identifying at-risk customers, allowing for proactive retention strategies.
3. **How do different machine learning models compare in their performance for churn prediction?** This includes evaluating and contrasting the accuracy, precision, recall, and F1-score of Logistic Regression, Random Forest, and XGBoost models to determine the most suitable model for this specific problem.

Key Performance Indicators (KPIs) and Evaluation Metrics:

- Accuracy
- Precision (for both 'No Churn' and 'Churn' classes)
- Recall (for both 'No Churn' and 'Churn' classes)
- F1-score (for both 'No Churn' and 'Churn' classes)
- Macro F1
- Weighted F1

What Success Looks Like

For this customer churn prediction project, success will be measured by the following criteria:

1. **Achieving Acceptable Prediction Accuracy**: The developed models (Logistic Regression, Random Forest, XGBoost) should demonstrate a minimum of 75% accuracy in predicting customer churn, with a strong emphasis on recall for the churn class (Churn=Yes) to minimize false negatives.
2. **Identifying Actionable Insights**: The exploratory data analysis (EDA) and feature importance analysis should reveal clear, interpretable patterns and drivers of churn. These insights should be actionable, guiding business strategies to reduce churn, such as identifying specific customer segments at high risk or pinpointing services/contract types that lead to higher churn rates.
3. **Developing a Deployable Model**: A robust and well-documented churn prediction model, preferably the best-performing one, should be developed and saved in a format suitable for deployment (e.g., `model.pkl`). This includes saving necessary pre-processing components like encoders. The project should culminate in a functional Streamlit application (`app.py`) that can take customer data as input and predict churn, demonstrating the practical utility and deployability of the solution.

What is the Impact?

Implementing this customer churn prediction system offers significant business value and benefits for the telecom company:

- **Reduced Customer Attrition and Improved Retention:** By identifying customers at high risk of churning *before* they leave, the company can proactively intervene with targeted retention strategies (e.g., special offers, personalized support, service upgrades). This directly leads to fewer lost customers and a higher overall customer retention rate.
- **Optimized Marketing and Resource Allocation:** The system allows for more precise targeting of marketing efforts. Instead of broad campaigns, resources can be focused on segments of customers who are both at risk of churning and responsive to retention efforts. This reduces marketing waste and maximizes the return on investment for retention campaigns.
- **Enhanced Customer Satisfaction:** Proactive engagement based on churn predictions can address customer pain points before they escalate. By understanding the factors driving churn (e.g., contract type, monthly charges, lack of tech support), the company can improve service quality and customer experience, leading to higher satisfaction and loyalty.
- **Increased Lifetime Value (LTV):** Retaining customers for longer periods directly increases their lifetime value. A customer who stays with the company for an additional year or more contributes significantly more revenue over time, making retention efforts highly profitable.
- **Data-Driven Decision Making:** The feature importance analysis from the models (e.g., Contract, Tenure, Monthly Charges, Internet Service Type, Online Security, Tech Support) provides actionable insights into what truly drives customer decisions. This allows the company to make strategic adjustments to its services, pricing, and customer support, fostering a more customer-centric business model.

Project Overview:

The main objective of this project was to predict customer churn in a telecom company. This involved analyzing various customer attributes to identify factors contributing to churn and building predictive models to proactively identify at-risk customers.

Dataset Used:

The dataset utilized for this project was "/content/WA_Fn-UseC_-Telco-Customer-Churn.csv", containing 7043 rows (customers) and 21 columns (features) with 'Churn' as the target variable.

Analysis Steps:

1. **Data Cleaning and Preprocessing:**
    - The `TotalCharges` column, initially an object type, was converted to a numeric (float) type. During this process, null values introduced by coercion were implicitly handled as `NaN`. (However, it was observed that no nulls were initially present based on `df.isna().sum()` and `df.info()` output, but `pd.to_numeric` can introduce them).
    - Checked for and confirmed no duplicate rows were present.

- No outliers were explicitly checked or handled.

2. **Exploratory Data Analysis (EDA) and Feature Engineering:**
    - **Correlation Analysis:** Explored the relationships between numerical features (`SeniorCitizen`, `tenure`, `MonthlyCharges`, `TotalCharges`). Key findings included a strong positive correlation between `tenure` and `TotalCharges` (0.83), and a moderate positive correlation between `MonthlyCharges` and `TotalCharges` (0.65).
    - **Churn Distribution:** Examined the overall distribution of churn, revealing an imbalanced dataset with more 'No' churn instances than 'Yes' churn.
    - **Impact Analysis:** Analyzed the impact of various features on churn:
        - `MonthlyCharges`: Customers who churned had higher average monthly charges. Further binned `MonthlyCharges` into custom categories (`18-40`, `41-60`, `61-80`, `81-100`, `101+`) to observe churn rates, noting higher churn in the `61-80` and `81-100` categories.
        - `gender`: No significant difference in churn rates between genders when considering `MonthlyCharges`.
        - `tenure`: Customers who churned had a significantly lower average tenure (approx. 18 months) compared to those who did not churn (approx. 37.5 months).
        - `Contract`: Longer contract terms were associated with lower average monthly charges and lower churn rates.
        - **Categorical Features:** Conducted `crosstab` analysis for `SeniorCitizen`, `Partner`, `Dependents`, `PhoneService`, `MultipleLines`, `InternetService`, `OnlineSecurity`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`, `Contract`, `PaperlessBilling`, and `PaymentMethod`. Key observations included higher churn among customers without online security, tech support, device protection, senior citizens, those with month-to-month contracts, and those using electronic checks.
    - **Feature Engineering:** Created a new categorical feature `MonthlyCharges_Category_Custom` by binning the `MonthlyCharges` column to analyze churn patterns within specific charge ranges.

Model Training and Evaluation:

Three different machine learning models were trained and evaluated for churn prediction:

1. **Random Forest Classifier:**
    - **Configuration:** `n_estimators=200`, `max_depth=10`, `class_weight='balanced'`, `random_state=42`.
    - **Evaluation Metrics:** Accuracy, Classification Report (precision, recall, f1-score for both classes), and Confusion Matrix.
    - **Performance:** Achieved an accuracy of 0.769, with a recall for churn (class 1) of 0.71 and precision of 0.55.

- **Feature Importance:** Identified `Contract`, `tenure`, `MonthlyCharges`, `OnlineSecurity`, `TechSupport`, and `InternetService` as key drivers of churn.

2. **Logistic Regression:**
   - **Configuration:** `max_iter=1000`, `class_weight='balanced'`, `random_state=42`.
   - **Evaluation Metrics:** Accuracy, Classification Report, and Confusion Matrix.
   - **Performance:** Achieved an accuracy of 0.736, with a recall for churn (class 1) of 0.80 and precision of 0.50.
   - **Feature Importance (Coefficients):** Highlighted `Contract` (negative coefficient, indicating retention for longer terms) and `PaperlessBilling` (positive coefficient, indicating churn driver) as most impactful.
3. **XGBoost Classifier:**
   - **Configuration:** `n_estimators=200`, `max_depth=5`, `learning_rate=0.1`, `scale_pos_weight` for class imbalance, `random_state=42`, `eval_metric='logloss'`.
   - **Evaluation Metrics:** Accuracy, Classification Report, and Confusion Matrix.
   - **Performance:** Achieved an accuracy of 0.748, with a recall for churn (class 1) of 0.76 and precision of 0.52.
   - **Feature Importance:** Similar to Random Forest, `Contract`, `InternetService`, `OnlineSecurity`, and `TechSupport` were important.

Model Comparison:

A comparison table and bar chart were created to visualize the performance of all three models across various metrics (Accuracy, Precision, Recall, F1-score for both 'No Churn' and 'Churn', Macro F1, and Weighted F1).

- **Overall:** Random Forest generally showed slightly better overall performance, especially in accuracy and weighted F1-score.
- **Churn Recall:** Logistic Regression had the highest recall for churn (0.80), followed closely by XGBoost (0.76) and Random Forest (0.71).
- **Business Insight:** While Random Forest or XGBoost might be preferred for production prediction due to slightly better overall accuracy and balanced F1-scores, Logistic Regression's coefficients offer clearer interpretability for explaining churn drivers to stakeholders.

Deployment Preparation:

The best-performing model (Random Forest) was selected for deployment. The model and encoders (for categorical features and target variable) were saved using `joblib` for future use in a Streamlit application (`app.py`), and a `requirements.txt` file was generated.

Model Recommendations:

Based on the comparative analysis, Random Forest and XGBoost models demonstrate superior predictive performance for identifying customer churn. Therefore, **Random Forest or XGBoost** are recommended for deployment in production environments where accurate churn prediction is paramount. Their higher accuracy and robust handling of complex relationships make them ideal for predictive tasks.

For explaining churn drivers to stakeholders and gaining interpretable insights, **Logistic Regression** is the preferred model. Its coefficients directly indicate the direction and magnitude of a feature's impact on churn, making it excellent for communicating 'why' a customer might churn.

Key Churn Drivers:

The analysis consistently highlighted several significant factors driving customer churn across all models:

- **Contract Type:** Month-to-month contracts are a very strong indicator of churn.
- **Tenure:** Customers with shorter tenures are significantly more likely to churn.
- **Monthly Charges:** Higher monthly charges are associated with increased churn, especially in the 61-100+ range.
- **Online Security:** Lack of online security services is a major churn driver.
- **Tech Support:** Absence of tech support services correlates with higher churn.
- **Internet Service Type:** Fiber optic internet service users show a higher propensity to churn compared to DSL users.
- **Paperless Billing:** Customers opted for paperless billing tend to churn more.
- **Senior Citizen:** Senior citizens have a higher churn rate.
- **Device Protection:** Lack of device protection also contributes to churn.

Actionable Retention Strategies:

To mitigate customer churn, the telecom company should implement targeted strategies focusing on the identified drivers:

1. **Target Month-to-Month Contract Customers:** Offer incentives (e.g., discounts, bundled services, loyalty bonuses) to encourage customers on month-to-month contracts to switch to longer-term contracts (one-year or two-year), thereby increasing their commitment.
2. **Early Tenure Engagement:** Implement proactive engagement programs for new customers, particularly within their first year, to ensure satisfaction and address any early issues that might lead to churn.
3. **Bundle Security & Support Services:** Actively promote and bundle online security and tech support services, especially for customers currently without them. Highlight the value and benefits of these services to improve retention.

4. **Address Fiber Optic Customer Satisfaction:** Investigate the reasons for higher churn among fiber optic users. This could involve improving service reliability, offering better customer support for technical issues specific to fiber optic, or re-evaluating pricing strategies for this segment.
5. **Review Paperless Billing Experience:** Understand why paperless billing is associated with higher churn. It might indicate a segment that is less engaged or more price-sensitive. Consider offering personalized communications or benefits to these customers.
6. **Senior Citizen Specific Programs:** Develop tailored plans, support services, or educational resources for senior citizens, addressing their specific needs and concerns to improve their loyalty.
7. **Optimize Pricing for High Monthly Charges:** For customers with high monthly charges, especially those in the $61-100+$ range, consider offering personalized plans, loyalty discounts, or value-added services to prevent them from seeking alternatives.
8. **Offer Device Protection Incentives:** Promote device protection plans more actively, perhaps by bundling them with other services or offering initial trial periods.