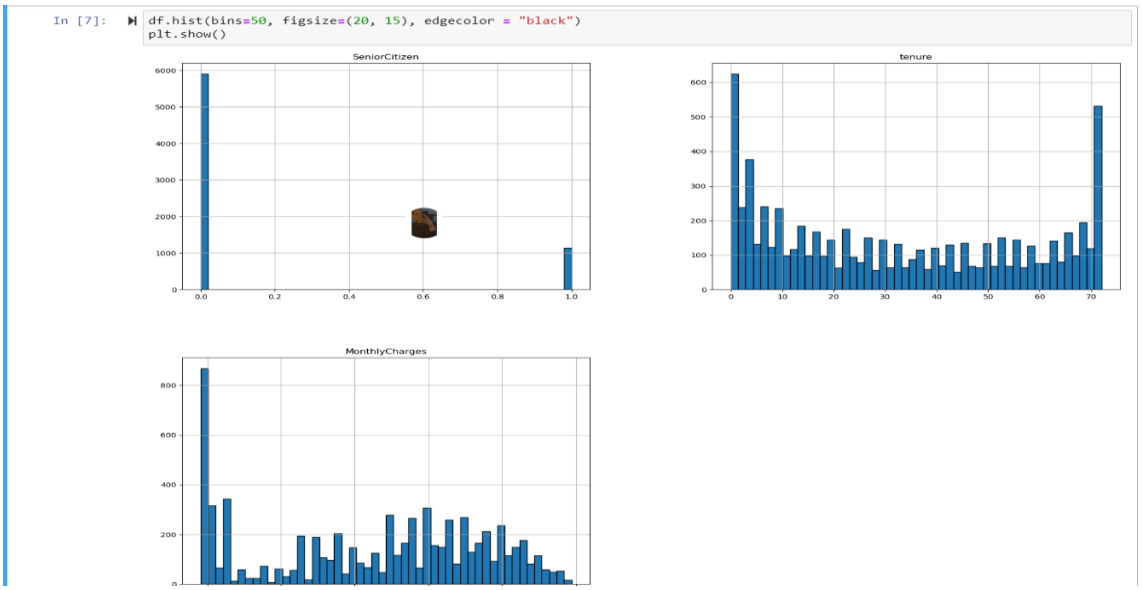
REPORT ON PREDICTING CUSTOMER CHURN IN A TELECOM COMPANY

PROBLEM DEFINITION:

The objective of this project is to develop a predictive model to identify customers likely to churn from TeleCom Inc. High churn rates negatively impact revenue and growth, so accurately predicting churn can help the company take proactive measures to improve customer retention.

DATA EXPLORATION AND PREPROCESSING

- **Dataset Structure**: The dataset consists of customer information, including demographic details, service usage patterns, and account information
- **Key Statistics and Feature Distribution**: Initial analysis reveals features such as customer tenure, monthly charges, total charges, and various categorical variables like gender, partner, dependents, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, and churn status.



- **Missing Values**: Detected in the `TotalCharges` column, which were handled by converting the column to numeric using `pd.to_numeric(errors='coerce')`. This method ensures that any nonnumeric values are turned into NaN, which can then be handled appropriately.

PREPROCESSING STEPS:

Handling Missing Values

Issue: In the TotalCharges column, there were missing values that needed to be addressed. The TotalCharges column was initially detected as an object type due to non-numeric entries.

Chosen Method:

1. **Conversion to Numeric:** Used pd.to_numeric(df['TotalCharges'], errors='coerce') to convert the column to numeric. This method coerces non-numeric values to NaN, ensuring that the column is interpreted correctly as numeric.

   Rationale: This approach allows for the identification and handling of non-numeric values systematically. By converting problematic entries to NaN, we can manage these missing values effectively.

2. **Imputation of Missing Values:** After conversion, missing values in the TotalCharges column were imputed using the median value of the column.

   Rationale: Imputing with the median is a robust approach because the median is less sensitive to outliers compared to the mean. This ensures that the imputed values do not skew the distribution of the TotalCharges column.

3. **Removing Irrelevant Features**: Removed `customerID` as it is unique for each customer and does not provide value for predicting churn. Also removed `Churn` column as it is the target variable.

4. **Encoding Categorical Variables**: Converted categorical variables using onehot encoding for nominal categories. Used label encoding for binary categorical variables.

5. **Scaling Numerical Features:** Applied standard scaling to ensure all numerical features are on a similar scale, aiding model stability.

FEATURE ENGINEERING:

- **Encoding and Scaling**: As discussed in preprocessing, categorical variables were encoded, and numerical features scaled.
- **Feature Selection:**
  Correlation Analysis: Performed to select features with high correlation to churn.
  Mutual Information: Assessed to identify features providing significant information about churn.

Justification:

The feature engineering process aimed to capture the essential patterns and relationships in the data that influence customer churn. The encoding and scaling steps ensured that the features were in a suitable format and scale for model training.

MODEL DEVELOPMENT: Algorithms Used

1. **Logistic Regression**

   Chosen for its simplicity and interpretability.

   Suitable for binary classification tasks like churn prediction.

2. **Decision Tree Classifier**

   Offers interpretability through decision rules.

   Capable of capturing nonlinear relationships in the data.

3. **Random Forest Classifier**

   An ensemble method that reduces overfitting by averaging multiple decision trees.

   Generally, provides better accuracy and stability compared to a single decision tree.

TRAINING AND EVALUATION:

Dataset Split: Split the dataset into training and testing sets (70% training, 30% testing) to evaluate model performance on unseen data.

Scaling: Standard scaling applied to ensure all features are on a similar scale, which helps in model convergence and stability.

Transformation and Fit: Transformed and fit the models on the training data, ensuring that scaling and encoding steps are applied consistently.

MODEL EVALUATION:

Performance Comparison:

- **Logistic Regression:** Basic model provided a benchmark for comparison. Achieved moderate accuracy and interpretability.
- **Decision Tree:** Higher interpretability but prone to overfitting. Provided insights into feature importance.
- **Random Forest:** Reduced overfitting with ensemble approach, performed the best among the three models.

Evaluation Metrics:

- **Accuracy:** Measures the overall correctness of the model predictions.
- **Precision:** Measures the correctness of positive predictions, i.e., the proportion of true positive predictions among all positive predictions.

- **Recall:** Measures the ability to capture actual positives, i.e., the proportion of true positives identified among all actual positives.
- **F1-Score:** Harmonic mean of precision and recall, providing a balance between the two metrics.

MODEL PERFORMANCE RESULTS:

Best Model: Random Forest Classifier, due to its balance between precision, recall, and overall accuracy.

```
# Collect and print evaluation results
print(classification_report(y_test,y_pred_lr))
```

```
              precision    recall  f1-score   support

          No       0.84      0.90      0.87      1041
         Yes       0.63      0.51      0.57       368

    accuracy                           0.79      1409
   macro avg       0.74      0.70      0.72      1409
weighted avg       0.78      0.79      0.79      1409
```

```
print(classification_report(y_test,y_pred_rf))
```

```
              precision    recall  f1-score   support

          No       0.83      0.89      0.86      1041
         Yes       0.61      0.47      0.53       368

    accuracy                           0.78      1409
   macro avg       0.72      0.68      0.69      1409
weighted avg       0.77      0.78      0.77      1409
```

```
print(classification_report(y_test,y_pred_dt))
```

```
              precision    recall  f1-score   support

          No       0.82      0.82      0.82      1041
         Yes       0.48      0.49      0.49       368

    accuracy                           0.73      1409
   macro avg       0.65      0.65      0.65      1409
weighted avg       0.73      0.73      0.73      1409
```

RECOMMENDATIONS:

1. **Targeted Promotions for New Customers:** Implement welcome programs and special promotions for new customers to enhance their initial experience and reduce the likelihood of early churn. This can include personalized onboarding, discounts, and proactive customer support.
2. **Value-based Pricing and Incentives:** Review and potentially revise the pricing strategy for high-charge customers. Offer value-based incentives, such as discounts, loyalty rewards, or bundled services, to provide perceived added value and encourage them to stay.
3. **Promoting Long-term Contracts:** Encourage customers to switch to longer-term contracts by offering attractive incentives like discounted rates, free additional services, or enhanced customer support. This can lock in customers for a longer period and reduce churn.
4. **Service Bundling and Cross-Selling:** Promote the use of multiple services through bundling offers. For example, customers with internet service could be offered discounted streaming services. Cross-selling additional services can make the overall package more attractive and harder to leave.
5. **Improving Payment Processes:** Investigate the reasons why certain payment methods are linked to higher churn and address those issues. Improving the convenience and reliability of the payment process can enhance overall customer satisfaction.
6. **Proactive Customer Support:** Use the churn predictions to identify high-risk customers and provide them with proactive support. This can include regular check-ins, personalized service reviews, and addressing any issues they might have before they decide to leave.

IMPLEMENTATION AND CONTINUOUS IMPROVEMENT

1. **Feedback Loop:** Continuously collect feedback from customers, especially those at risk of churning. Use this feedback to refine and enhance the predictive models and retention strategies.
2. **Data-Driven Decisions:** Regularly update the models with new data to ensure they remain accurate and relevant. Use the insights from the models to make informed, data-driven decisions about customer retention strategies.
3. **Performance Monitoring:** Track the effectiveness of the implemented strategies through key performance indicators (KPIs) such as churn rates, customer satisfaction scores, and average customer tenure. Adjust strategies as needed based on these metrics.
4. **Service Improvement:** Identify and address common issues faced by customers likely to churn, such as service interruptions or billing disputes.

By leveraging the insights from the predictive models, TeleCom Inc. can implement targeted and effective retention strategies to reduce customer churn, improve customer satisfaction, and ultimately enhance overall business performance.