Customer Churn Analysis & Prediction

Name : Zeyad Ahmed Mostafa

mail: ziada00700@gmail.com

Phone : +201200249877

Data Exploration &Analysis

1. Dataset Overview

- ### 1.1 Dataset Description

Dataset Size: The dataset contains 3,333 rows (observations) and 20 columns (features).

Target Variable: The target variable is Churn, a binary feature indicating whether a customer has churned (True or False).

Features:
Categorical columns >> ['International plan', 'Voice mail plan', 'Churn'] <<
Numerical columns >> ['Account length', 'Number vmail messages', 'Total day minutes', 'Total day calls', 'Total day charge', 'Total eve minutes', 'Total eve calls', 'Total eve charge', 'Total night minutes', 'Total night calls', 'Total night charge', 'Total intl minutes', 'Total intl calls', 'Total intl charge', 'Customer service calls'] <<

- ### 1.2 Data Types

Numerical Features:
Integer: Account length, Area code, Number vmail messages, Total day calls, Total eve calls, Total night calls, Total intl calls, Customer service calls.
Float: Total day minutes, Total day charge, Total eve minutes, Total eve charge, Total night minutes, Total night charge, Total intl minutes, Total intl charge.
Categorical Features:
Object: State, International plan, Voice mail plan.
Boolean: Churn.

- ### 1.3 Missing Values & Duplicated

No Missing Values & Duplicated: The dataset is complete, with no missing values in any of the columns. This simplifies preprocessing, as no imputation or removal of missing data is required.

- ### 1.4 Descriptive Statistics

Numerical Features:

	count	mean	std	min	25%	50%	75%	max
Account length	3333.000000	101.064806	39.822106	1.000000	74.000000	101.000000	127.000000	243.000000
Area code	3333.000000	437.182418	42.371290	408.000000	408.000000	415.000000	510.000000	510.000000
Number vmail messages	3333.000000	8.099010	13.688365	0.000000	0.000000	0.000000	20.000000	51.000000
Total day minutes	3333.000000	179.775098	54.467389	0.000000	143.700000	179.400000	216.400000	350.800000
Total day calls	3333.000000	100.435644	20.069084	0.000000	87.000000	101.000000	114.000000	165.000000
Total day charge	3333.000000	30.562307	9.259435	0.000000	24.430000	30.500000	36.790000	59.640000
Total eve minutes	3333.000000	200.980348	50.713844	0.000000	166.600000	201.400000	235.300000	363.700000
Total eve calls	3333.000000	100.114311	19.922625	0.000000	87.000000	100.000000	114.000000	170.000000
Total eve charge	3333.000000	17.083540	4.310668	0.000000	14.160000	17.120000	20.000000	30.910000
Total night minutes	3333.000000	200.872037	50.573847	23.200000	167.000000	201.200000	235.300000	395.000000
Total night calls	3333.000000	100.107711	19.568609	33.000000	87.000000	100.000000	113.000000	175.000000
Total night charge	3333.000000	9.039325	2.275873	1.040000	7.520000	9.050000	10.590000	17.770000
Total intl minutes	3333.000000	10.237294	2.791840	0.000000	8.500000	10.300000	12.100000	20.000000
Total intl calls	3333.000000	4.479448	2.461214	0.000000	3.000000	4.000000	6.000000	20.000000
Total intl charge	3333.000000	2.764581	0.753773	0.000000	2.300000	2.780000	3.270000	5.400000
Customer service calls	3333.000000	1.562856	1.315491	0.000000	1.000000	1.000000	2.000000	9.000000

Categorical Feature:

	count	unique	top	freq
State	3333	51	WV	106
International plan	3333	2	No	3010
Voice mail plan	3333	2	No	2411
Churn	3333	2	False	2850

2- Data Analysis & Visualization with insights:

Three key functions were implemented for analysis:

1. value_counts: Calculates unique values' count and percentage distribution in a specified column, used for categorical features like International plan and State.
2. calculate_churn_rates: Computes churn and existing customer rates for categorical columns, highlighting categories with the highest churn rates.
3. add_custom_stats:This function calculates minimum, maximum, and average statistics for specified columns and merges them with a value counts DataFrame. It also styles the output with bar plots for better visualization.

```
The International plan Value counts
```

	International plan	Count	Percentage (%)
0	No	3010	90.31
1	Yes	323	9.69

```
The International plan Value counts by Target(Attrition_Flag)
```

Churn	International plan	False	True	Total	churn Rate (%)	Existing Customer Rate (%)
0	No	2664	346	3010	11.495017	88.504983
1	Yes	186	137	323	42.414861	57.585139

```
The Voice mail plan Value counts
```

	Voice mail plan	Count	Percentage (%)
0	No	2411	72.34
1	Yes	922	27.66

```
The Voice mail plan Value counts by Target(Attrition_Flag)
```

Churn	Voice mail plan	False	True	Total	churn Rate (%)	Existing Customer Rate (%)
0	No	2008	403	2411	16.715056	83.284944
1	Yes	842	80	922	8.676790	91.323210

```
The Churn Value counts
```

	Churn	Count	Percentage (%)
0	False	2850	85.51
1	True	483	14.49

Analyze distribution across target categories

	Churn	Count	Percentage (%)	min_Total eve calls	max_Total eve calls	avg_Total eve calls
0	False	2850	85.51%	0.00	170.00	100.04
1	True	483	14.49%	48.00	168.00	100.56

	Churn	Count	Percentage (%)	min_Total eve charge	max_Total eve charge	avg_Total eve charge
0	False	2850	85.51%	0.00	30.75	16.92
1	True	483	14.49%	6.03	30.91	18.05

	Churn	Count	Percentage (%)	min_Total night minutes	max_Total night minutes	avg_Total night minutes
0	False	2850	85.51%	23.20	395.00	200.13
1	True	483	14.49%	47.40	354.90	205.23

	Churn	Count	Percentage (%)	min_Total night calls	max_Total night calls	avg_Total night calls
0	False	2850	85.51%	33.00	175.00	100.06
1	True	483	14.49%	49.00	158.00	100.40

	Churn	Count	Percentage (%)	min_Total night charge	max_Total night charge	avg_Total night charge
0	False	2850	85.51%	1.04	17.77	9.01
1	True	483	14.49%	2.13	15.97	9.24

	Churn	Count	Percentage (%)	min_Total intl minutes	max_Total intl minutes	avg_Total intl minutes
0	False	2850	85.51%	0.00	18.90	10.16
1	True	483	14.49%	2.00	20.00	10.70

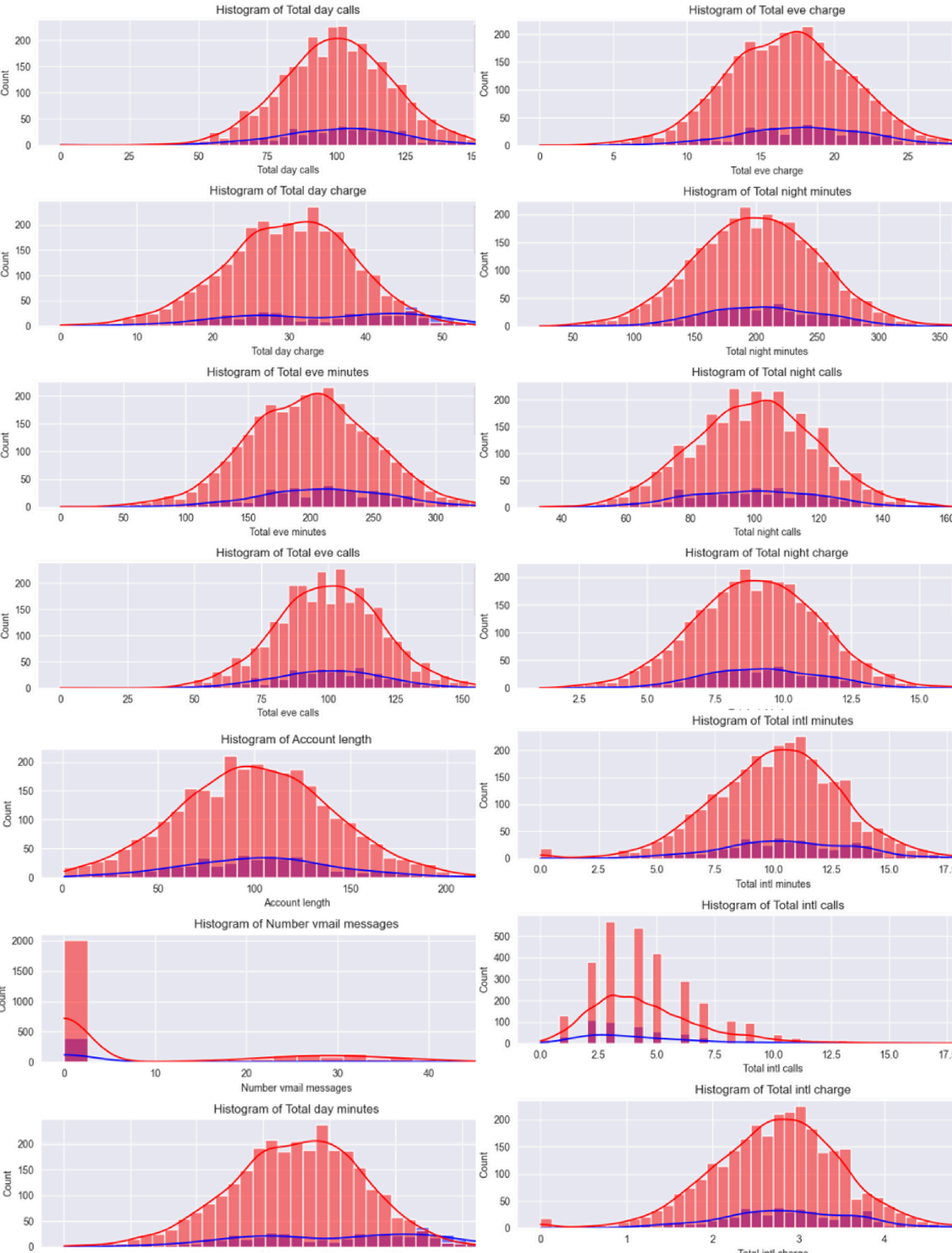
	Churn	Count	Percentage (%)	min_Account length	max_Account length	avg_Account length
0	False	2850	85.51%	1.00	243.00	100.79
1	True	483	14.49%	1.00	225.00	102.66

	Churn	Count	Percentage (%)	min_Number vmail messages	max_Number vmail messages	avg_Number vmail messages
0	False	2850	85.51%	0.00	51.00	8.60
1	True	483	14.49%	0.00	48.00	5.12

Insights

1. Churn Rate: The overall churn rate is 14.49%, with 85.51% of customers remaining active.
2. Evening Calls and Charges:
   - The average number of evening calls is similar for both churned and non-churned customers (around 100).
   - Churned customers have slightly higher average evening charges (18.05) compared to non-churned customers (16.92).
3. Night Minutes and Charges:
   - Churned customers have slightly higher average night minutes (205.23) and charges (9.24) compared to non-churned customers (200.13 and 9.01, respectively).
4. International Minutes:
   - Churned customers have slightly higher average international minutes (10.70) compared to non-churned customers (10.16).
5. Account Length:
   - The average account length is slightly higher for churned customers (102.66) compared to non-churned customers (100.79), indicating that account length might not be a significant factor in churn.
6. Voicemail Messages:
   - Non-churned customers have a higher average number of voicemail messages (8.60) compared to churned customers (5.12), suggesting that customers who use voicemail more frequently are less likely to churn.
7. Day Minutes and Charges:
   - Churned customers have significantly higher average day minutes (206.91) and charges (35.18) compared to non-churned customers (175.18 and 29.78, respectively). This indicates that higher daytime usage might be associated with higher churn rates.
8. Day Calls:
   - The average number of day calls is similar for both churned and non-churned customers (around 100), indicating that the number of calls might not be a significant factor in churn.
9. Evening Minutes:
   - Churned customers have higher average evening minutes (212.41) compared to non-churned customers (199.04), suggesting that higher evening usage might be associated with higher churn rates.

distribution of numerical features:



distribution of Categorical features

Numerical Features man by Churn Status

Preprocessing & Model building

Preprocessing:

1. Encoding:
    - The target variable Churn was encoded (False: 0, True: 1).
    - Categorical features (International plan and Voice mail plan) were label-encoded.
2. Train-Test Split:
    - The dataset was split into features (X) and target (y), and further divided into training (80%) and testing (20%) sets.
3. Handling Class Imbalance:
    - A class imbalance in Churn was observed using a count plot.
    - SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the training data.
4. Log Transformation:
    - Log transformation was applied to skewed numerical columns (Number vmail messages, Total intl calls, Customer service calls) to normalize their distributions.
5. Scaling:
    - Numerical features were standardized using StandardScaler to ensure consistent scaling across the dataset.

Model Building and Evaluation

1. XGBoost Model:
    - An XGBoost classifier was trained without hyperparameter tuning.
    - Achieved high performance:
        - Accuracy: 96.4%
        - Recall: 80.6%
        - F1-Score: 86.2%
        - Precision: 92.5%
2. Other Models with Hyperparameter Tuning:
    - Multiple models (Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, AdaBoost, Gradient Boosting) were trained using GridSearchCV with recall as the scoring metric.
    - Top Performers:
        - Gradient Boosting:
            - Accuracy: 96.2%
            - Recall: 78.4%
            - F1-Score: 85.3%
            - Precision: 93.5%
        - Random Forest:
            - Accuracy: 94.1%
            - Recall: 66.67%
            - F1-Score: 76%
            - Precision: 88.5%
3. Model Comparison:
    - XGBoost outperformed other models in terms of accuracy, recall, and F1-score.
    - Gradient Boosting and Random Forest also showed strong performance, with high accuracy and balanced precision-recall trade-offs.

Key Insights

- Feature Importance:
    - XGBoost's feature importance plot highlighted the most influential features for predicting churn.

	Model	Accuracy	Recall	F1-Score	Precision	Best Parameters
0	XGBoost	0.964018	0.806452	0.862069	0.925926	{Without grid-search}
7	Gradient Boosting	0.962519	0.784946	0.853801	0.935897	{'learning_rate': 0.1, 'max_depth': 7, 'min_sa...
5	Random Forest	0.941529	0.666667	0.760736	0.885714	{'criterion': 'gini', 'max_depth': 20, 'min_sa...
4	Decision Tree Classifier	0.893553	0.774194	0.669767	0.590164	{'criterion': 'entropy', 'max_depth': 20, 'min...
6	AdaBoost	0.887556	0.569892	0.585635	0.602273	{'learning_rate': 1, 'n_estimators': 100}
3	K-Nearest Neighbors	0.805097	0.559140	0.444444	0.368794	{'algorithm': 'auto', 'n_neighbors': 3, 'weigh...
1	Logistic Regression	0.743628	0.634409	0.408304	0.301020	{'C': 0.1, 'max_iter': 100, 'penalty': 'l2'}
2	Naive Bayes	0.566717	0.741935	0.323185	0.206587	{}

Confusion Matrix