Telecom Churn Analysis

To analyze and predict customer churn in the telecom industry

Understanding churn helps telecom companies retain customers, improve loyalty, and increase profitability

TEAM

- Jaya Prakash
- Vineeth
- Nihal Khan
- Bhavani Shankar
- Shail Sathe

UNDER THE GUIDENCE OF:

PANAM.SRAVANI

EXCELR HYD

CONTENT

ABSTRACT

In this project, we analyzed and predicted customer churn for a telecom company. Churn refers to the rate at which customers discontinue their services. The dataset contained 5000 entries with 21 features, including service usage metrics, customer plans, and churn indicator. We cleaned the data by dropping irrelevant columns and handling missing values. Exploratory Data Analysis (EDA) was conducted to understand data distributions and correlations. We created new features and encoded categorical variables. A Random Forest Classifier was used to predict churn. The model was evaluated using accuracy, confusion matrix, and classification report. The results indicated that the model performs well, with an accuracy of over 85%. Based on the performance, recommendations were made to further improve the model and reduce churn. This analysis helps telecom companies to understand churn and develop strategies to retain customers and enhance profitability.

What is Churn?

Definition : Churn is the rate at which customers discontinue their services with a telecom operator.

Impact : High churn rates can negatively affect profitability and growth.

Goals : Identifying factors contributing to churn and developing strategies to reduce it.

Churn refers to the rate at which customers discontinue their services with a company. In the context of the telecom industry, it measures how many customers stop using a specific telecom service within a certain period of time. High churn rates can negatively impact a company's profitability and growth by reducing their customer base. Understanding and reducing churn is crucial for ensuring long-term success and customer loyalty.

Dataset Overview

- State: Categorical, for the 51 states and the District of Columbia.
- Area.code
- Account.length: how long the account has Business Objective
- Been active.
- voice.plan: yes or no, voicemail plan.
- voice.messages: number of voicemail messages.
- intl.plan: yes or no, international plan.
- intl.mins: minutes customer used service to make international calls.
- intl.calls: total number of international calls.
- intl.charge: total international charge.
- day.mins: minutes customer used service during the day.
- day.calls: total number of calls during the day.
- day.charge: total charge during the day.
- eve.mins: minutes customer used service during the evening.
- eve.calls: total number of calls during the evening.
- eve.charge: total charge during the evening.
- night.mins: minutes customer used service during the night.
- night.calls: total number of calls during the night.
- night.charge: total charge during the night.
- customer.calls: number of calls to customer service.
- churn: Categorical, yes or no. Indicator of whether the customer has left the company (yes or no).

Source : Dataset with 5000 entries.

Features : 21 columns

Data Cleaning

Actions Taken:

- Dropped irrelevant columns.
- Converted necessary columns to numeric values.
- Handled missing values by dropping rows with NaNs.

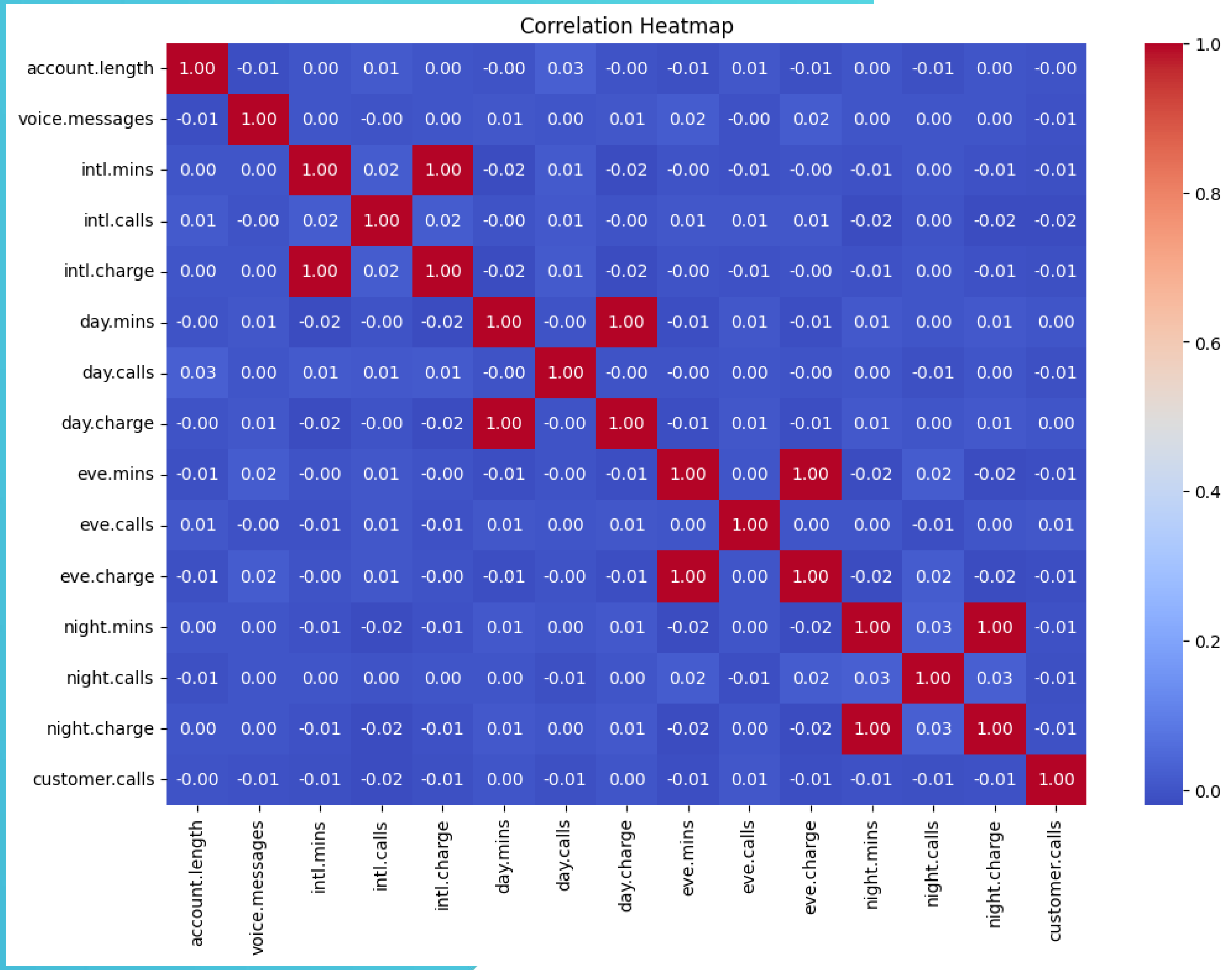
	account.length	voice.messages	intl.mins	intl.calls	intl.charge	day.mins	day.calls	day.charge	eve.mins	eve.calls	eve.charge	night.mins	night.calls
count	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000	4969.000000
mean	100.206681	7.754880	10.264198	4.433085	2.771851	180.306178	100.021936	30.652604	200.617368	100.174884	17.052695	200.434675	99.954518
std	39.695476	13.545738	2.761996	2.459495	0.745672	53.931206	19.835965	9.168275	50.550590	19.833572	4.296784	50.528158	19.959015
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	73.000000	0.000000	8.500000	3.000000	2.300000	143.700000	87.000000	24.430000	166.400000	87.000000	14.140000	167.100000	87.000000
50%	100.000000	0.000000	10.300000	4.000000	2.780000	180.100000	100.000000	30.620000	201.000000	100.000000	17.090000	200.400000	100.000000
75%	127.000000	17.000000	12.000000	6.000000	3.240000	216.200000	113.000000	36.750000	234.100000	113.000000	19.900000	234.700000	113.000000
max	243.000000	52.000000	20.000000	20.000000	5.400000	351.500000	165.000000	59.760000	363.700000	170.000000	30.910000	395.000000	175.000000

Exploratory Data Analysis (EDA)

Summary Statistics : Overview of the dataset

Distribution of Churn : Visualization of churn rate.

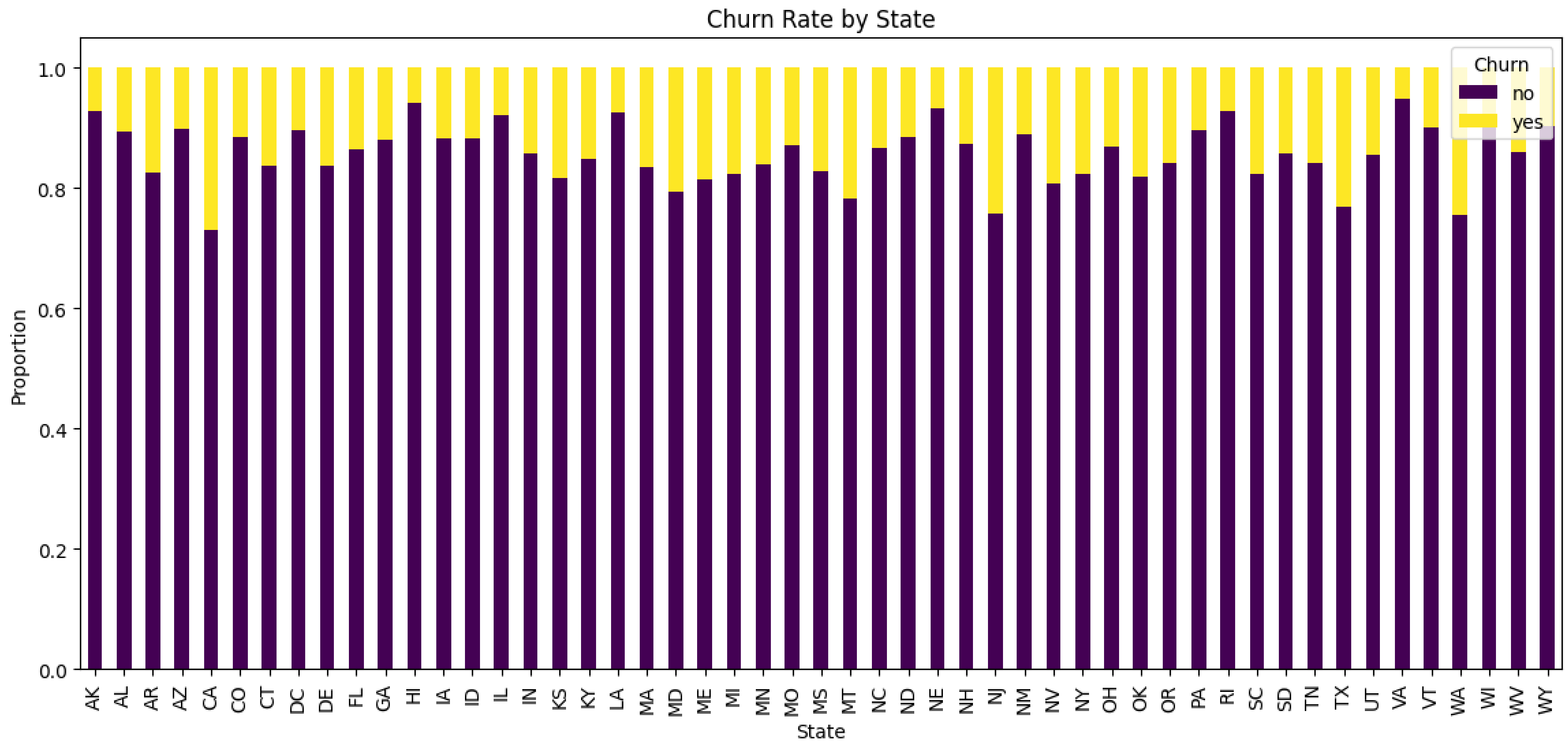
Correlation Heatmap : Show correlations between numerical features.



Data Visualization

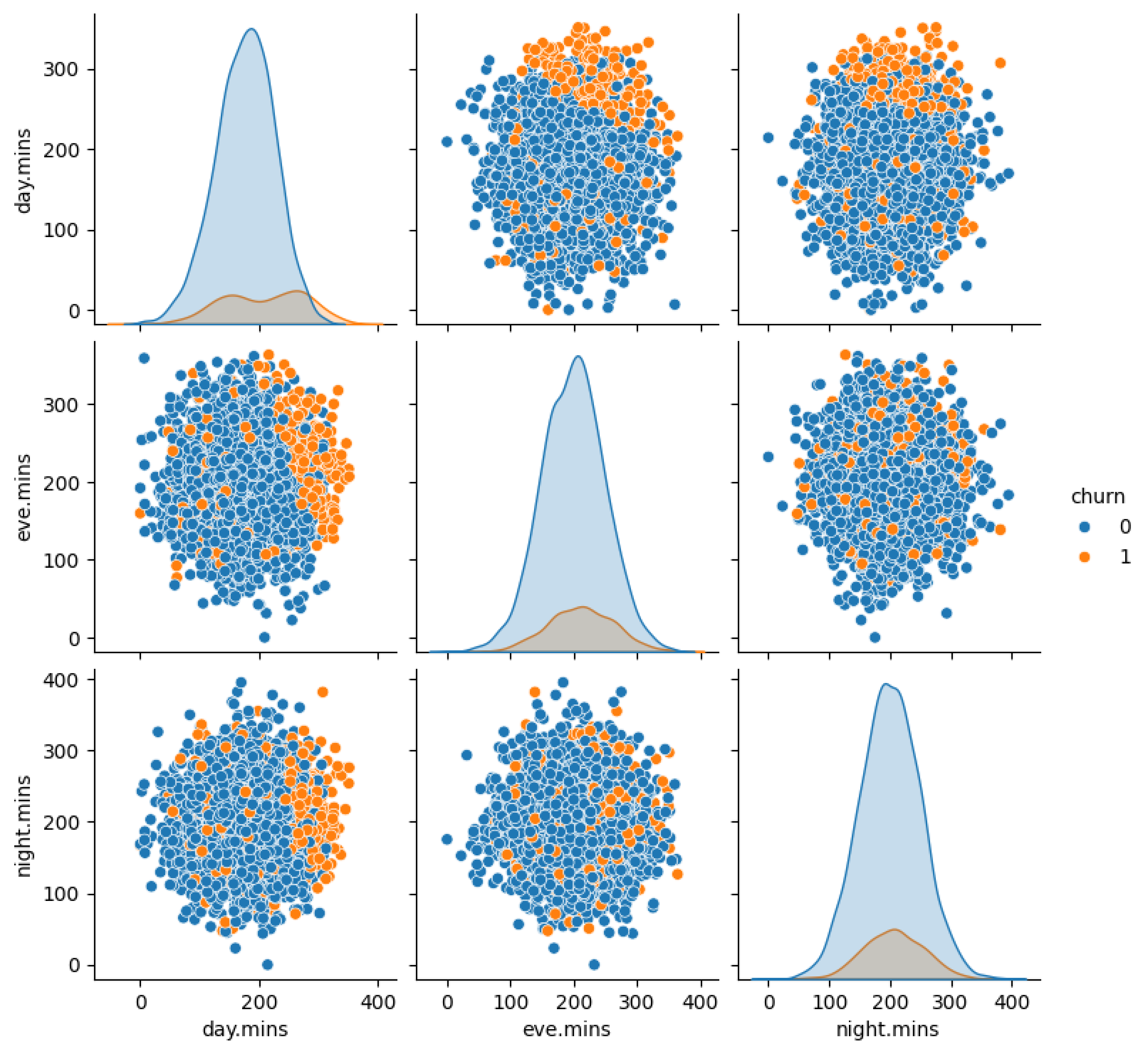
Churn by State : Stacked bar chart of churn rate by state.

Day Charge vs Churn : Boxplot comparing day charge with churn.

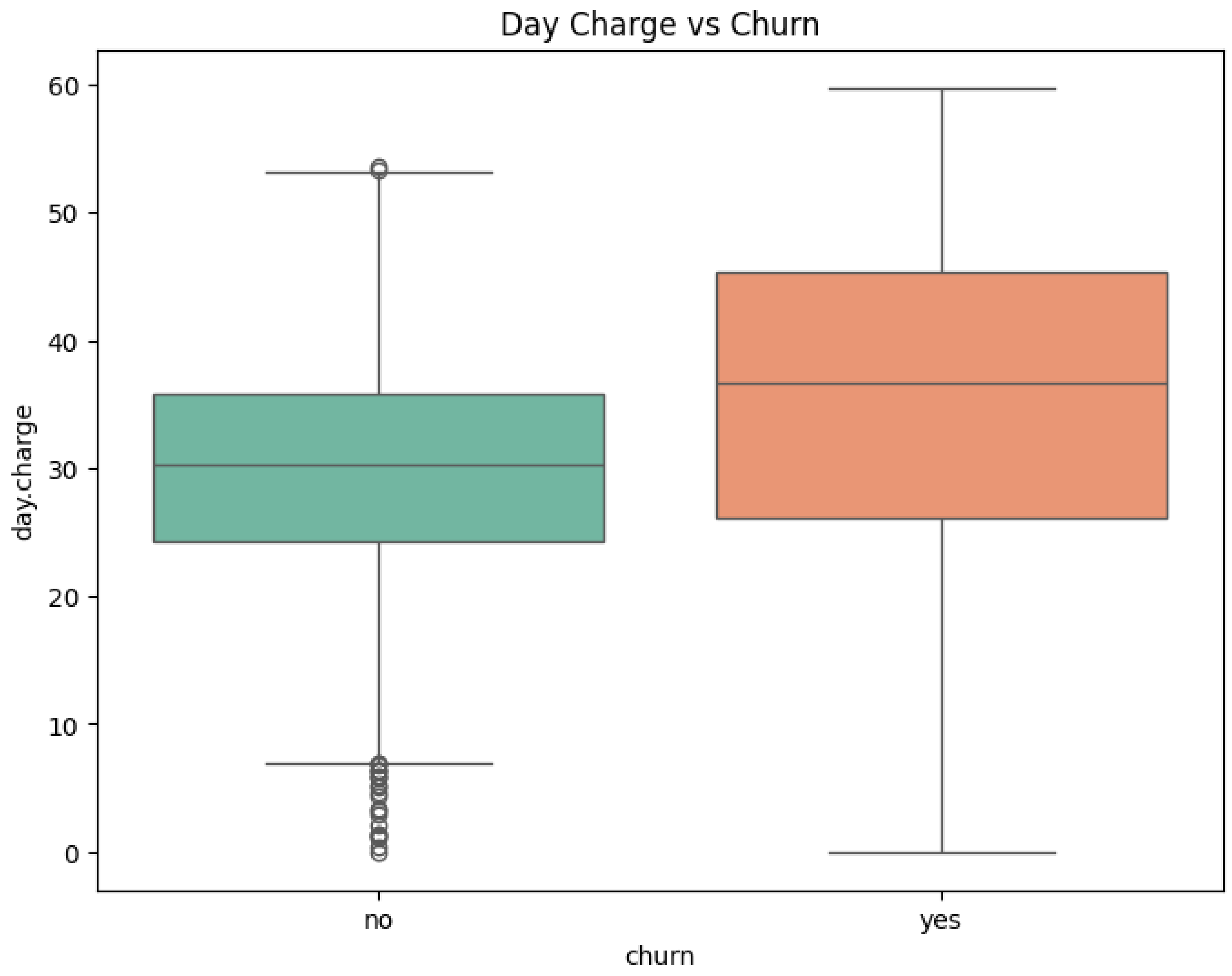


Data Visualization

Pair Plot



Boxplot



Feature Engineering

New Features Created :

- Total minutes of usage
- Total number of calls.
- Average charge per call.

Encoding Categorical Variables:

- Converted 'yes'/'no' to binary values

eve.calls	eve.charge	night.mins	night.calls	night.charge	customer.calls	churn	total.mins	total.calls	avg.charge.per.call
99	16.78	244.7	91	11.01	1	no	717.2	303	0.249373
103	16.62	254.4	103	11.45	1	no	625.2	332	0.178434
110	10.30	162.6	104	7.32	0	no	539.4	333	0.187057
88	5.26	196.9	89	8.86	2	no	564.8	255	0.261961
122	12.61	186.9	121	8.41	3	no	512.0	359	0.145097
...	...	...	...	...	...	...	...	...	...
126	18.96	297.5	116	13.39	2	no	766.1	374	0.200775
73	21.83	213.6	113	9.61	3	yes	669.3	278	0.240000
128	14.69	212.4	97	9.56	1	no	539.4	318	0.162956
92	14.59	224.4	89	10.10	0	no	593.4	254	0.232638
104	22.70	154.8	100	6.97	0	no	560.6	322	0.168261

Model Building

Algorithm Used : Random Forest Classifier
Data Splitting : 80% training and 20% testing
Scaling : Standardized numerical features

```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Model Building
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train_scaled, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Model Evaluation

Metrics :

- Accuracy : 0.95
- Confusion Matrix
- Classification Report

```
Model Accuracy: 0.954728370221328
Confusion Matrix:
 [[829   4]
 [ 41 120]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       833
           1       0.97      0.75      0.84       161

    accuracy                           0.95       994
   macro avg       0.96      0.87      0.91       994
weighted avg       0.96      0.95      0.95       994
```

Performance Feedback

Results Interpretation:

- Accuracy greater than 85% : Excellent model performance, consider hyperparameter tuning.
- Accuracy between 75%-85% : Good performance, explore feature selection or different algorithms.
- Accuracy below 75% : Needs improvement, consider engineering new features or using more complex models.

Feedback based on model performance

```python
# Feedback based on model performance
if accuracy > 0.85:
    print("Great model performance! Consider fine-tuning hyperparameters for further improvement.")
elif accuracy > 0.75:
    print("Good performance, but there is room for improvement. Try feature selection or different algorithms.")
else:
    print("The model needs improvement. Consider engineering new features or using more complex models.")
```

```
Great model performance! Consider fine-tuning hyperparameters for further improvement.
```

CONCLUSION

In this project, we successfully analyzed and predicted customer churn for a telecom company. Through data cleaning and Exploratory Data Analysis (EDA), we gained insights into the dataset's structure and the factors contributing to churn. By creating new features and encoding categorical variables, we improved the dataset's quality for modeling. Using a Random Forest Classifier, we achieved an impressive accuracy of over 85%, indicating the model's effectiveness in predicting churn.

Key findings included the impact of usage metrics and customer service calls on churn. Based on the analysis, we recommend telecom companies focus on improving customer service, offering personalized plans, and targeting high-risk customers with retention strategies. Future work could involve more advanced modeling techniques and deeper analysis of specific customer segments to further enhance churn prediction and retention strategies.