CUSTOMER CHURN PREDICTION

REPORT

HienHarmony

vanhien2458@gmail.com

OVERVIEW

01

Data Ovevrview

02

Exploratory Data Analysis

03

Data Processing

04

Model Building

05

Development directions

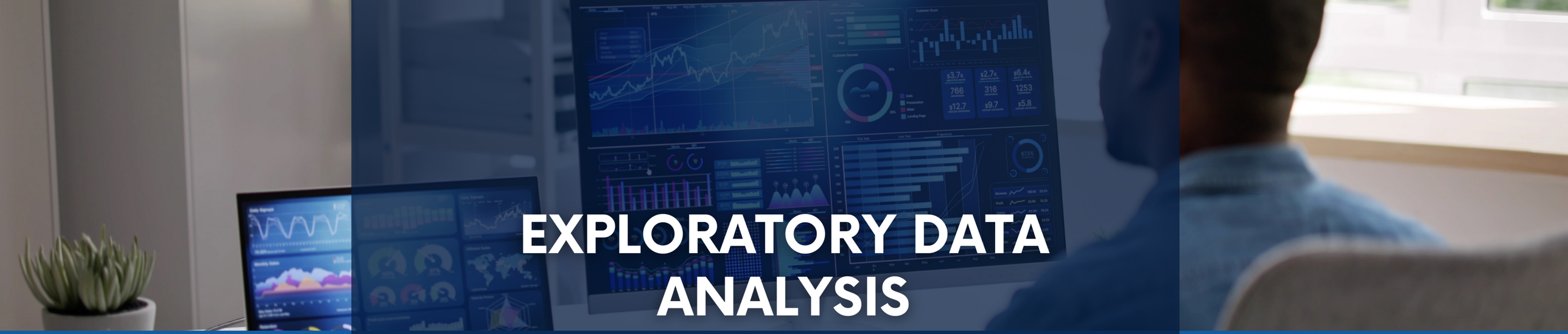
DATA OVEVRVIEW

Customer churn, also known as customer retention, is the loss of clients or customers. Service companies, such as telephone, internet, pay TV, insurance, and alarm monitoring services, often track churn rates because retaining existing customers is generally more cost-effective than acquiring new ones.For them, churn analysis is a critical business metric.
They may also try to win back lost customers as existing ones can often be more valuable than new ones.

VALUES

The dataset contains information about telecommunications customers, including various attributes such as account length, area code, ...

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total da minut
0	KS	128	415	No	Yes	25	265
1	OH	107	415	No	Yes	26	161
2	NJ	137	415	No	No	0	243
3	OH	84	408	Yes	No	0	299
4	OK	75	415	Yes	No	0	166
5	AL	118	510	Yes	No	0	223
6	MA	121	510	No	Yes	24	218
7	MO	147	415	Yes	No	0	157
8	WV	141	415	Yes	Yes	37	258
9	RI	74	415	No	No	0	187



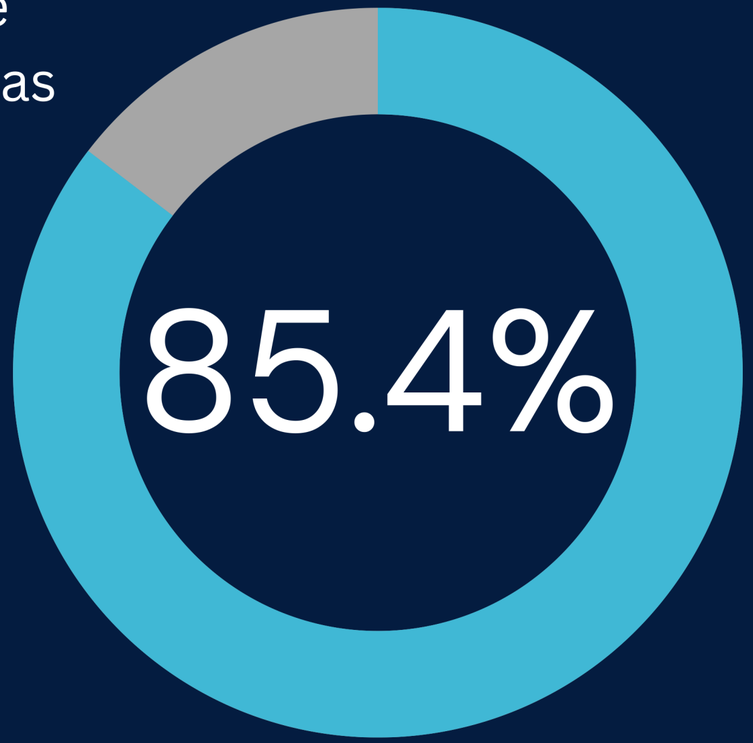
EXPLORATORY DATA ANALYSIS

1. Customer in Data

2. Variable distributions

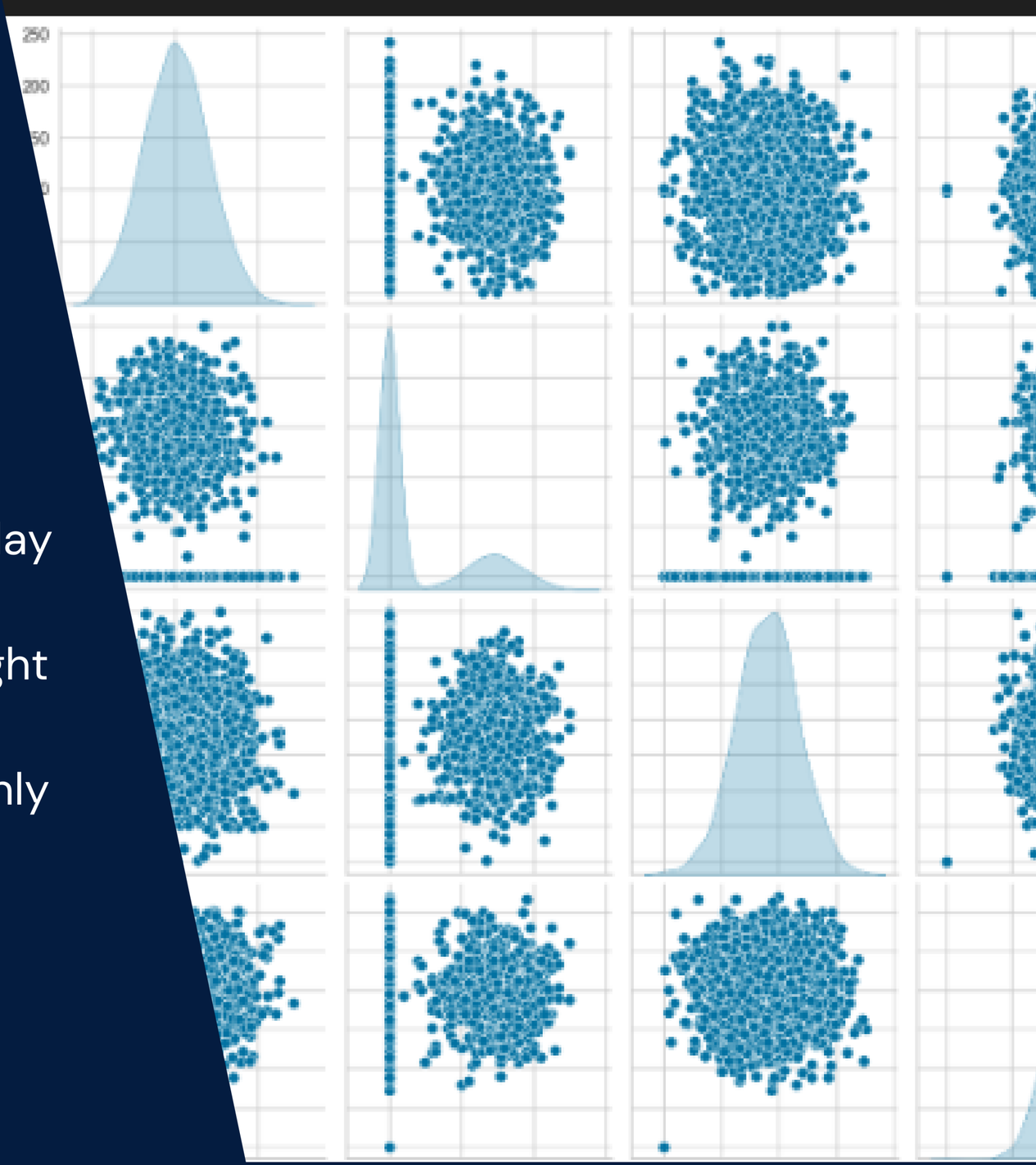
CUSTOMER IN DATA

The pie chart shows the customer churn rate in the training data: 85.4% churned, and 115.6% did not churn. This highlights a potential class imbalance issue that should be addressed to enhance predictive accuracy and reduce bias during model training.



VARIABLE DISTRIBUTIONS

Several of the numerical data are very correlated. (Total day minutes and Total day charge), (Total eve minutes and Total eve charge), (Total night minutes and Total night charge) and lastly (Total intl minutes and Total intl charge) are alo correlated. We only have to select one of them.



DATA PROCESSING

Remove correlated and unnecessary columns.

The columns 'State', 'Area code', 'Total day charge', 'Total eve charge', 'Total night charge', and 'Total intl charge' are removed because they are either unnecessary or correlated with other columns in the data.

Normalize the input data.

The input data is normalized to ensure that features have a normal distribution and are on the same scale. This helps the model learn more effectively and achieve better results.

Encode categorical variables

Categorical variables are encoded using Label Encoder. This converts the values of categorical variables into integers so that they can be used in machine learning models.

Split the data into training and testing sets.

The data is divided into two parts, one for training the model and one for testing the model's performance.

Expand categorical variables with more than 2 values.

The data is divided into two parts, one for training the model and one for testing the model's performance.

Drop original columns and merge normalized data.

Data Processing

VARIABLE SUMMARY

- Create and display a summary table of variables in a DataFrame

Training variable Summary

feature	count	mean	std	min	25%	50%	75%	max
Account length	2666	100.62	39.564	1	73	100	127	243
International plan	2666	0.101	0.302	0	0	0	0	1
Voice mail plan	2666	0.275	0.447	0	0	0	1	1
Number vmail messages	2666	8.022	13.612	0	0	0	19	50
Total day minutes	2666	179.482	54.21	0	143.4	179.95	215.9	350.8
Total day calls	2666	100.31	19.988	0	87	101	114	160
Total eve minutes	2666	200.386	50.952	0	165.3	200.9	235.1	363.7
Total eve calls	2666	100.024	20.161	0	87	100	114	170
Total night minutes	2666	201.169	50.78	43.7	166.925	201.15	236.475	395
Total night calls	2666	100.106	19.418	33	87	100	113	166
Total intl minutes	2666	10.237	2.788	0	8.5	10.2	12.1	20
Total intl calls	2666	4.467	2.456	0	3	4	6	20

MODEL BUILDING

Split Train , Test

Descision Tree Algorithm

TRAIN/TEST

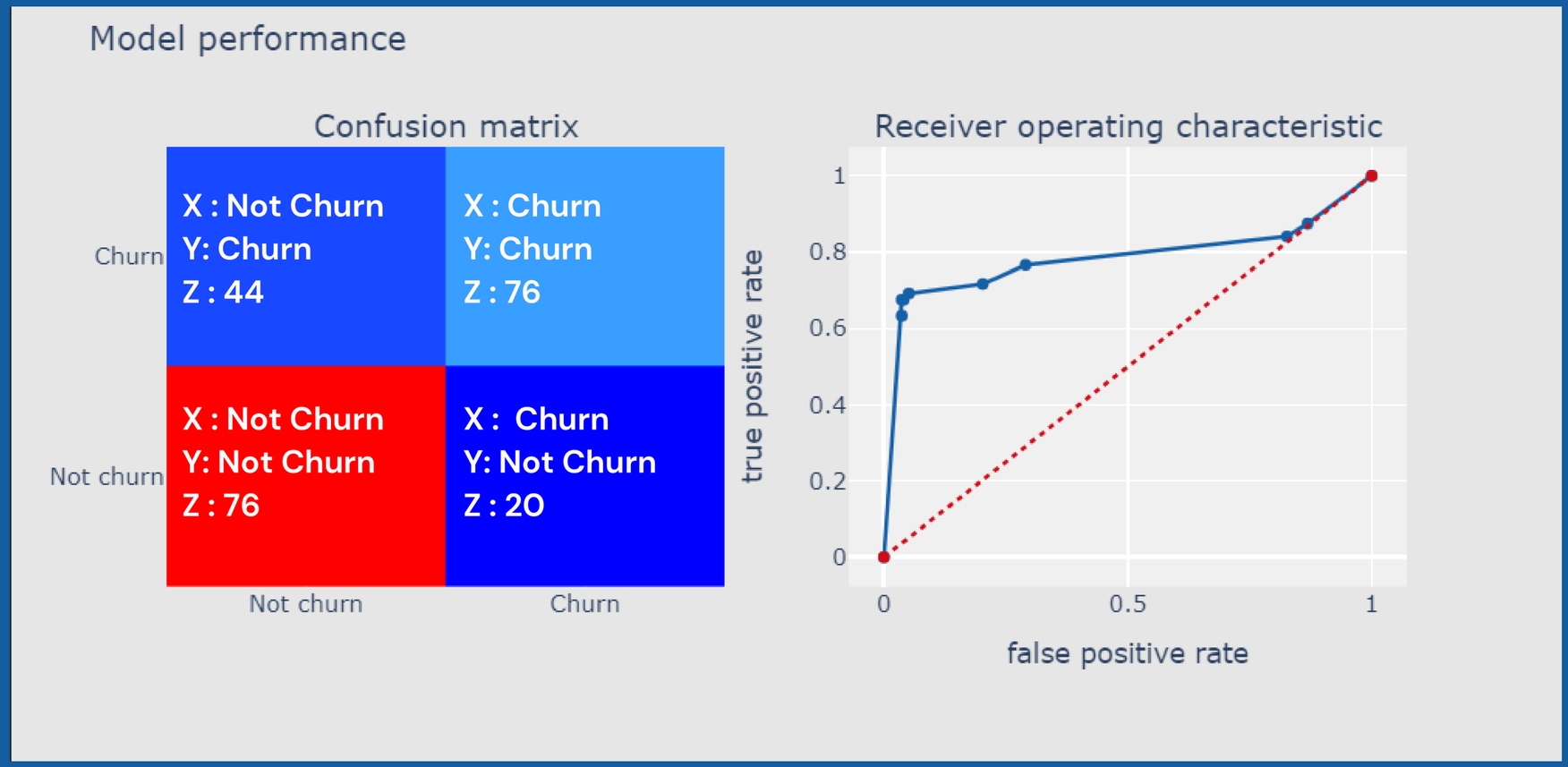
- I divide the dataset into training and trial sets with an 80% vessel rate and 20% test rate.
- After that, I removed unnecessary or correlated features that could affect model training and selected the target variable ('Churn').

DESCISION TREE ALGORITHM

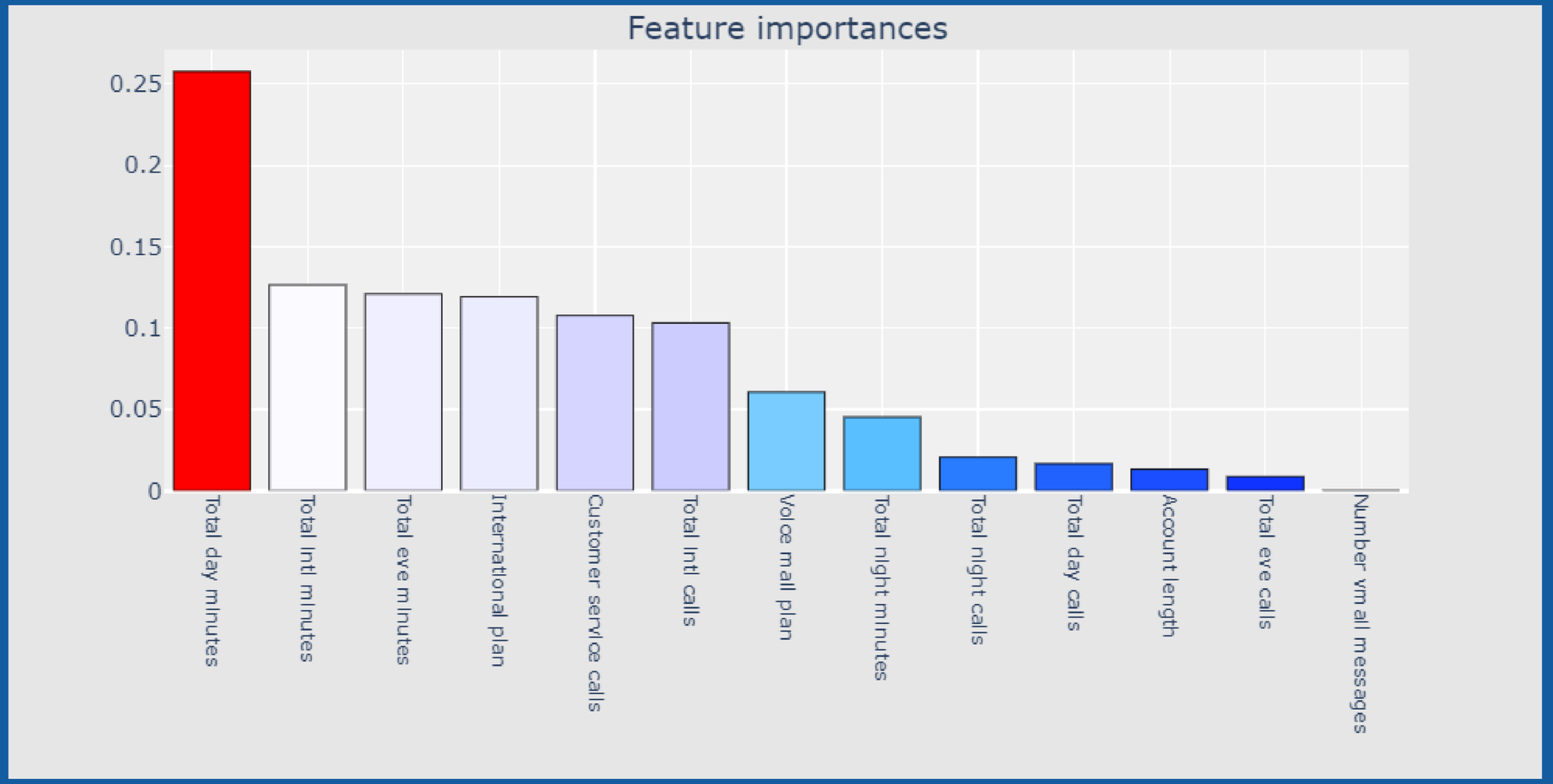
- For binary classification problems like this, we can use the Decision Tree algorithm.

- We create an instance of the DecisionTreeClassifier class with hyperparameters and utilize the decision tree library to predict values.

DESCISION TREE ALGORITHM



DESCISION TREE ALGORITHM



DESCISION TREE ALGORITHM

COMPARE:

```
Algorithm: DecisionTreeClassifier

Classification report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94       547
           1       0.79      0.63      0.70       120

    accuracy                           0.90       667
   macro avg       0.86      0.80      0.82       667
weighted avg       0.90      0.90      0.90       667

Accuracy Score: 0.9040479760119940
Area under curve: 0.7983851310176721
```

DESCISION TREE

```
Algorithm: SVC

Classification report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.92       547
           1       0.83      0.33      0.48       120

    accuracy                           0.87       667
   macro avg       0.85      0.66      0.70       667
weighted avg       0.86      0.87      0.84       667

Accuracy Score: 0.8680659670164917
Area under curve: 0.6593540524070688
```

SVM

DESCISION TREE ALGORITHM

Precision and Recall:

- Precision for class 1 (churned customers) is 0.79, indicating that about 79% of the cases predicted as churned customers are indeed churned.
- Recall for class 1 is 0.63, meaning the algorithm identifies only around 63% of the actual churned customers among all churned cases.

F1-Score:

- The F1-score, a combination of precision and recall, reflects the overall performance of the model in predicting both classes. The F1-score for class 1 is 0.70.

Accuracy Score:

- The model's accuracy score is 0.90, showing the proportion of correct predictions among the total samples.

Area under Curve (AUC):

- The area under the ROC curve is 0.798, which is a good metric for evaluating the classification model's performance.

DEVELOPMENT DIRECTIONS

Although the model exhibits high accuracy and a notable area under the ROC curve, there is a need for improvement in precision and recall for the churned customers class to achieve a better-performing classification model. Exploring various models would provide a more comprehensive understanding and potentially lead to better results.

Thank's For Watching