PHASE 3 PROJECT PRESENTATION

PROBLEM STATEMENT

- **Customer Churn Analysis in Telecom**

Customer retention is essential for long-term business success in the competitive telecom industry.

Churn occurs when users discontinue services, impacting revenue and growth.

Identifying key factors driving churn enables proactive strategies to enhance customer satisfaction.

This analysis uses customer data to build a predictive model for identifying high-risk customers.

BUSINESS OBJECTIVES

1. Predict churn probability: Develop a model to classify customers as likely to churn or likely to stay based on their usage behavior.
2. Identify key churn indicators: Determine which factors (e.g., high call charges, frequent customer service calls, etc.) contribute the most to churn.

DATA CLEANING AND UNDERSTANDING

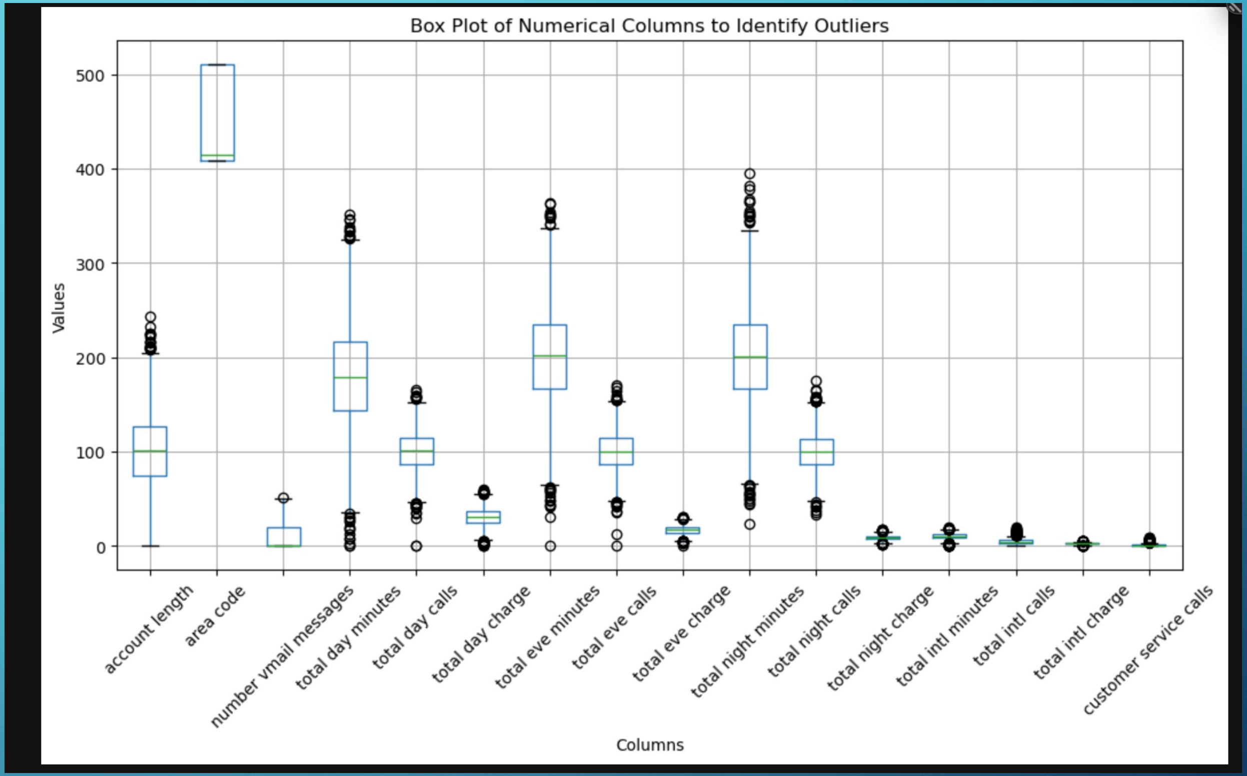
- After carefull analysis of the data, none of the rows seem to have null or duplicated values
- I proceeded to look at outlier values

ANALYSIS OF BOXPLOTS

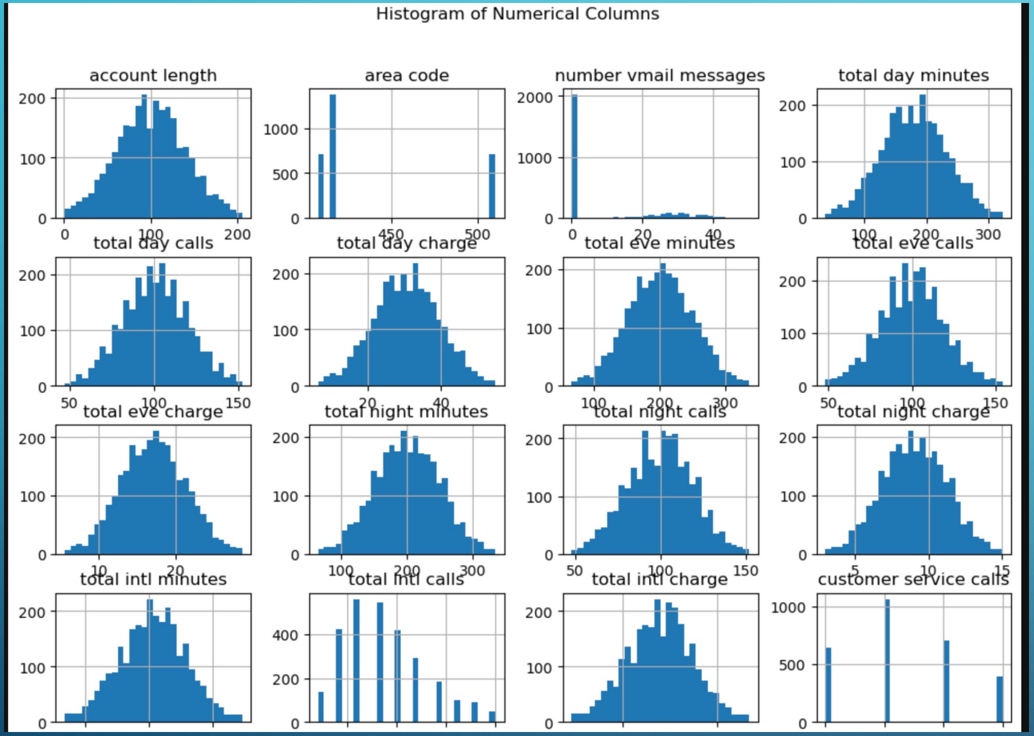
Box plot to my right shows how the data is distributed

I dropped all rows that contained outliers as they would affect the performance of my model later

I then proceeded to doing EDA

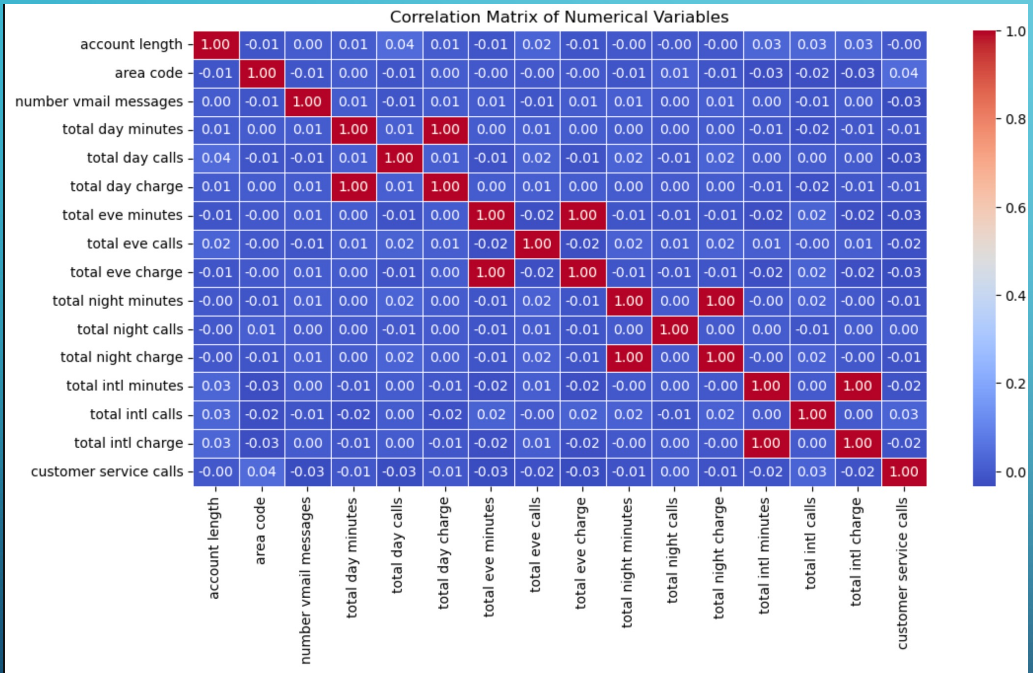


EDA



- First I drew a histogram showing how all numerical data is distributed
- From the diagram most of the columns appear to have there data as normally ditsributed

CORRELATION MATRIX



Correlation Matrix of Numerical Variables

	account length	area code	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
account length	1.00	-0.01	0.00	0.01	0.04	0.01	-0.01	0.02	-0.01	-0.00	-0.00	-0.00	0.03	0.03	0.03	-0.00
area code	-0.01	1.00	-0.01	0.00	-0.01	0.00	-0.00	-0.00	-0.00	-0.01	0.01	-0.01	-0.03	-0.02	-0.03	0.04
number vmail messages	0.00	-0.01	1.00	0.01	-0.01	0.01	0.01	-0.01	0.01	0.01	0.00	0.01	0.00	-0.01	0.00	-0.03
total day minutes	0.01	0.00	0.01	1.00	0.01	1.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	-0.02	-0.01	-0.01
total day calls	0.04	-0.01	-0.01	0.01	1.00	0.01	-0.01	0.02	-0.01	0.02	-0.01	0.02	0.00	0.00	0.00	-0.03
total day charge	0.01	0.00	0.01	1.00	0.01	1.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	-0.02	-0.01	-0.01
total eve minutes	-0.01	-0.00	0.01	0.00	-0.01	0.00	1.00	-0.02	1.00	-0.01	-0.01	-0.01	-0.02	0.02	-0.02	-0.03
total eve calls	0.02	-0.00	-0.01	0.01	0.02	0.01	-0.02	1.00	-0.02	0.02	0.01	0.02	0.01	-0.00	0.01	-0.02
total eve charge	-0.01	-0.00	0.01	0.00	-0.01	0.00	1.00	-0.02	1.00	-0.01	-0.01	-0.01	-0.02	0.02	-0.02	-0.03
total night minutes	-0.00	-0.01	0.01	0.00	0.02	0.00	-0.01	0.02	-0.01	1.00	0.00	1.00	-0.00	0.02	-0.00	-0.01
total night calls	-0.00	0.01	0.00	0.00	-0.01	0.00	-0.01	0.01	-0.01	0.00	1.00	0.00	0.00	-0.01	0.00	0.00
total night charge	-0.00	-0.01	0.01	0.00	0.02	0.00	-0.01	0.02	-0.01	1.00	0.00	1.00	-0.00	0.02	-0.00	-0.01
total intl minutes	0.03	-0.03	0.00	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.00	0.00	-0.00	1.00	0.00	1.00	-0.02
total intl calls	0.03	-0.02	-0.01	-0.02	0.00	-0.02	0.02	-0.00	0.02	0.02	-0.01	0.02	0.00	1.00	0.00	0.03
total intl charge	0.03	-0.03	0.00	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.00	0.00	-0.00	1.00	0.00	1.00	-0.02
customer service calls	-0.00	0.04	-0.03	-0.01	-0.03	-0.01	-0.03	-0.02	-0.03	-0.01	0.00	-0.01	-0.02	0.03	-0.02	1.00

- **Perfect Correlation (1.00):** Certain features have a direct linear relationship.
- **Key Correlated Pairs:**

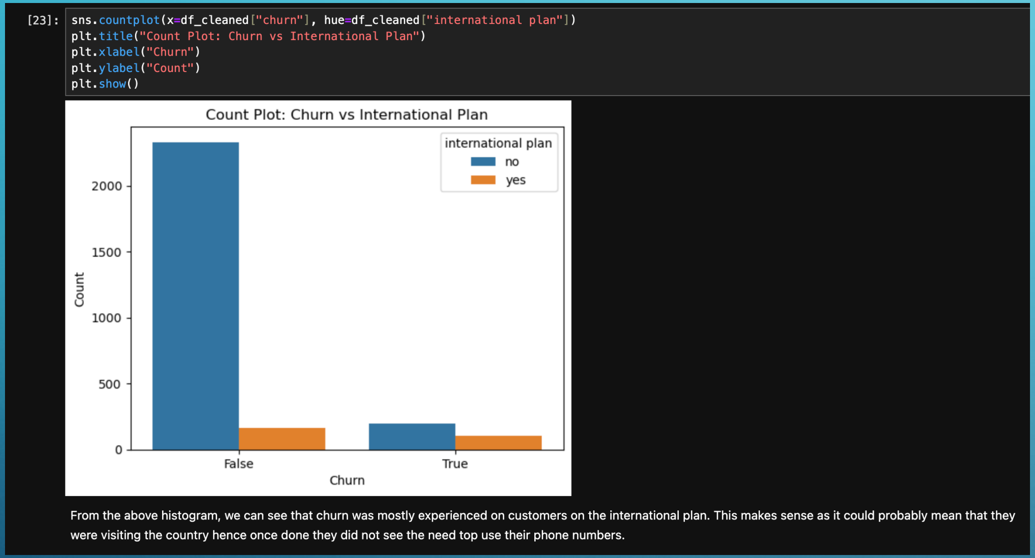
  Total Day Minutes ↔ Total Day Charge

  Total Evening Minutes ↔ Total Evening Charge

  Total Night Minutes ↔ Total Night Charge

  Total International Minutes ↔ Total International Charge
- **Impact:** These redundant features will be removed to simplify the model without losing information.

RELATIONSHIP BETWEEN CHURN AND INTERNATIONAL PLAN



```
[23]: sns.countplot(x=df_cleaned["churn"], hue=df_cleaned["international plan"])
      plt.title("Count Plot: Churn vs International Plan")
      plt.xlabel("Churn")
      plt.ylabel("Count")
      plt.show()
```

From the above histogram, we can see that churn was mostly experienced on customers on the international plan. This makes sense as it could probably mean that they were visiting the country hence once done they did not see the need top use their phone numbers.

- From the above histogram, we can see that churn was mostly experienced on customers on the international plan.
- Probably because they were visiting the country hence once done they did not see the need continue using their phone numbers.

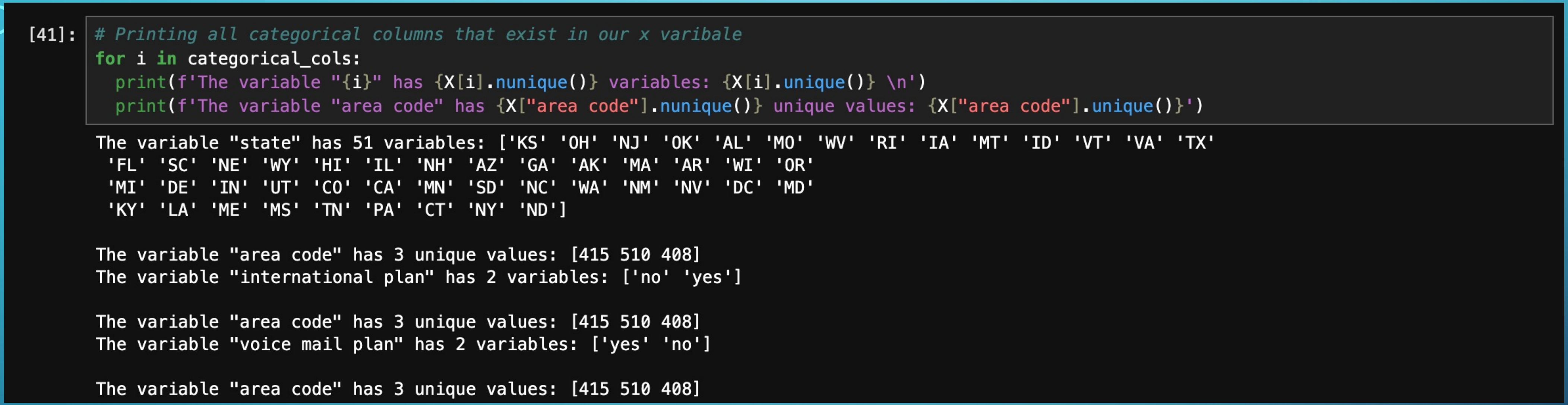
PRE PROCESSING

- After EDA and data understanding, I proceed to Pre-processing of the data
- I Defined X (features) and Y (target)
- For my Y(target) variable I defined it as churn(i.e the number of customers who left the telecommunications company)

ENCODING

- Upon analysis of the target variable(churn), it only had two valuesi.e true and false
- I then encoded the Target varibale(churn) into numerical format i.e. 1s and 0s to enable me perform the analysis
- In addition, I identified the categorical columns in my feature variables(X)

ENCODING(CONTD)



```
[41]: # Printing all categorical columns that exist in our x varibale
      for i in categorical_cols:
        print(f'The variable "{i}" has {X[i].nunique()} variables: {X[i].unique()} \n')
        print(f'The variable "area code" has {X["area code"].nunique()} unique values: {X["area code"].unique()}')
```

```
The variable "state" has 51 variables: ['KS' 'OH' 'NJ' 'OK' 'AL' 'MO' 'WV' 'RI' 'IA' 'MT' 'ID' 'VT' 'VA' 'TX'
 'FL' 'SC' 'NE' 'WY' 'HI' 'IL' 'NH' 'AZ' 'GA' 'AK' 'MA' 'AR' 'WI' 'OR'
 'MI' 'DE' 'IN' 'UT' 'CO' 'CA' 'MN' 'SD' 'NC' 'WA' 'NM' 'NV' 'DC' 'MD'
 'KY' 'LA' 'ME' 'MS' 'TN' 'PA' 'CT' 'NY' 'ND']

The variable "area code" has 3 unique values: [415 510 408]
The variable "international plan" has 2 variables: ['no' 'yes']

The variable "area code" has 3 unique values: [415 510 408]
The variable "voice mail plan" has 2 variables: ['yes' 'no']

The variable "area code" has 3 unique values: [415 510 408]
```

Only 4 columns had categorical data i.e. state, areaCode, Voice mail and international plan

ENCODING AND SCALING (CONTD)

- I encoded categorical columns and converted the values into numerical for analysis purposes
- I then proceeded to scale data using standard scale
- Scaling data ensures that numerical features are on a similar scale, which can improve model performance and training stability.
- Finally I split the data into test and training data

MODELING

- I then proceed to create several Classification models to gauge the performance and identify the best model

MODEL	ACCURACY	TARGET	PRECISON	RECALL	F1-SCORE
Logistic Regression	0.91	Churn	0.92	0.98	0.95
		Non-churn	0.76	0.40	0.52
Tuned Logistic Regression	0.76	Churn	0.96	0.76	0.85
		Non-churn	0.32	0.77	0.45
Random Forest	0.91	Churn	0.91	1.00	0.95
		Non-churn	0.92	0.33	0.48
Tuned Random Forest	0.93	Churn	0.93	1.00	0.96
		Non-churn	0.95	0.50	0.65
Decision Tree	0.92	Churn	0.95	0.96	0.95
		Non-churn	0.69	0.64	0.67

MODELING

More models continued

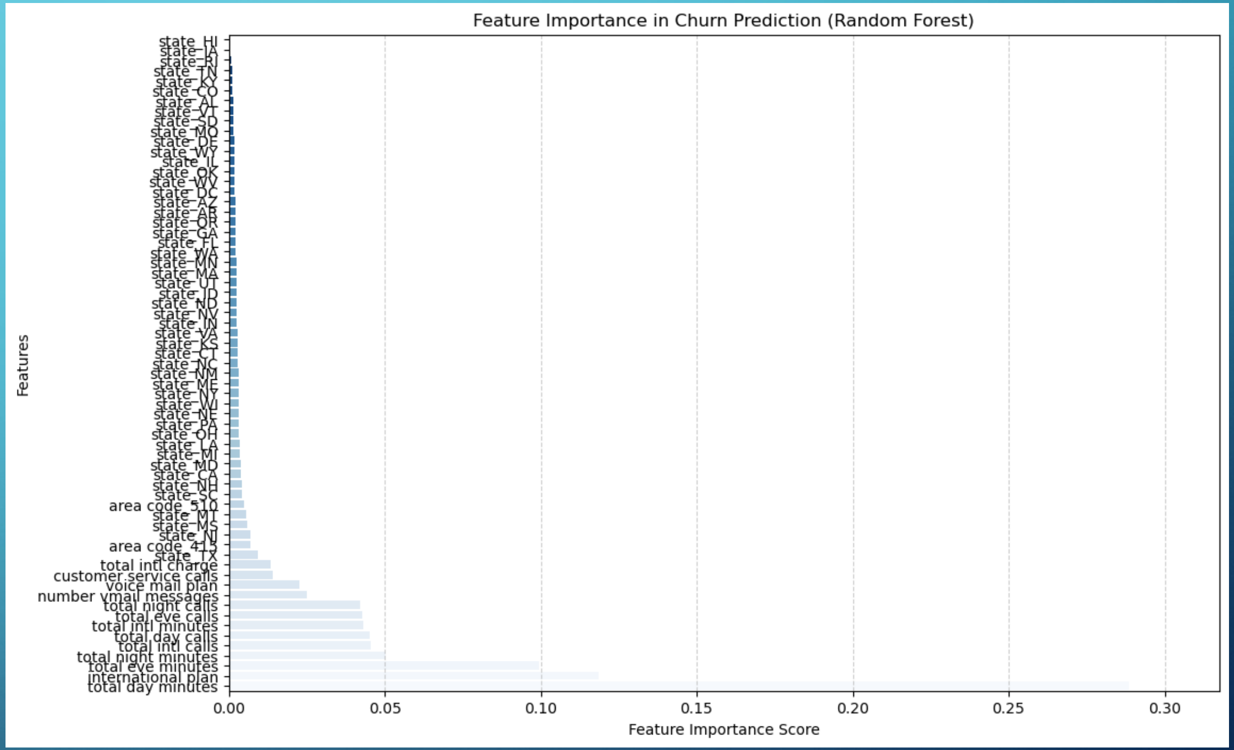
MODEL	ACCURACY	TARGET	PRECISON	RECALL	F1-SCORE
Decision Tree Tuned	0.93	Churn	0.96	0.96	0.96
		Non-churn	0.74	0.73	0.73
Support Vector Machine	0.90	Churn	0.94	0.95	0.94
		Non-churn	0.60	0.54	0.57
Support Vector Machine Tuned	0.89	Churn	0.94	0.94	0.94
		Non-churn	0.56	0.57	0.57
KNN classifier	0.87	Churn	0.88	0.99	0.93
		Non-churn	0.20	0.01	0.03

CHOSING BEST MODEL

- Best model would be the tuned Random Forest. This is because:
- It has the Best balance between precision (0.95) and recall (0.50) for churned customers
- It has the Highest accuracy (93%)
- And finally it Performs better than Decision Tree, SVM, and Logistic Regression on f1-score

FEATURES AFFECTING CHURN

From the barplot we can see that total day minutes affected churn the most followed by international plan



CONCLUSION

In conclusion, the tuned random forest was the best model to perform the predictions as it had highest precison of 0.95 and highest churn recall(0.5) out of all models

- Precision (0.95 for churned customers)

Out of all customers the model predicted as "churned," 95% were actually churned This means we have very few false positives (customers mistakenly identified as churned)

- Recall (0.50 for churned customers)

Out of all actual churned customers, only 50% were correctly identified This means some churned customers were missed (false negatives)

Top 3 Features which majorly affect the churn are total day minutes, international plan and total eveing minutes.