ST1 Capstone project

Predicting Tip Amounts in a Restaurant Setting

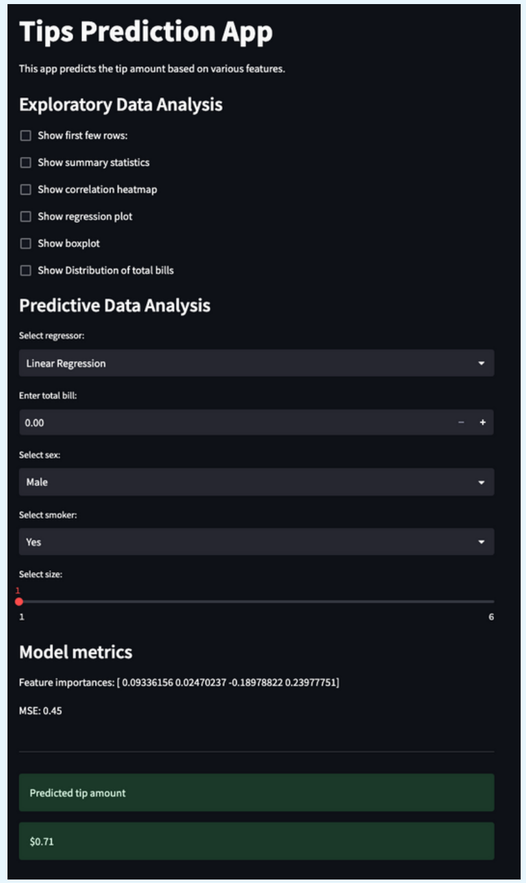
Eh Ser Tin
U3257929
4483_BRUCE_SEM-1_ON-CAMPUS
Software Technology 1 Thu 15:30

Table and contents

Predicting Tip Amounts in a Restaurant Setting

Problem Description:

How different factors like total bill amount, gender, smoker smoker and size affect the tip amount in a restaurant?



- DATASET: KAGGLE WAITER'S RECORDED TIPS (244 OBSERVAT. `1IONS, 7 VARIABLES)

- Exploratory Data Analysis (EDA) to identify patterns and trends

- Predictive Data Analysis (PDA) using Linear Regression and Random Forest Regressor

- Implementation & Deployment: Streamlit web app for real-time tip predictions and visualizations

Dataset Details

Dataset Description

total_bill	float64
tip	float64
sex	object
smoker	object
day	object
time	object
size	int64

5 rows | 5 rows x 7 columns pd.Dataframe | CSV

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

DATASET SOURCE: KAGGLE: TIPS.CSV

DATASET SIZE: 244 OBSERVATIONS

VARIABLES:

- TIP AMOUNT
- TOTAL BILL
- SEX
- SMOKER STATUS
- DAY
- TIME
- PARTY SIZE

EDA (Exploratory Data Analysis)
Summary statistics

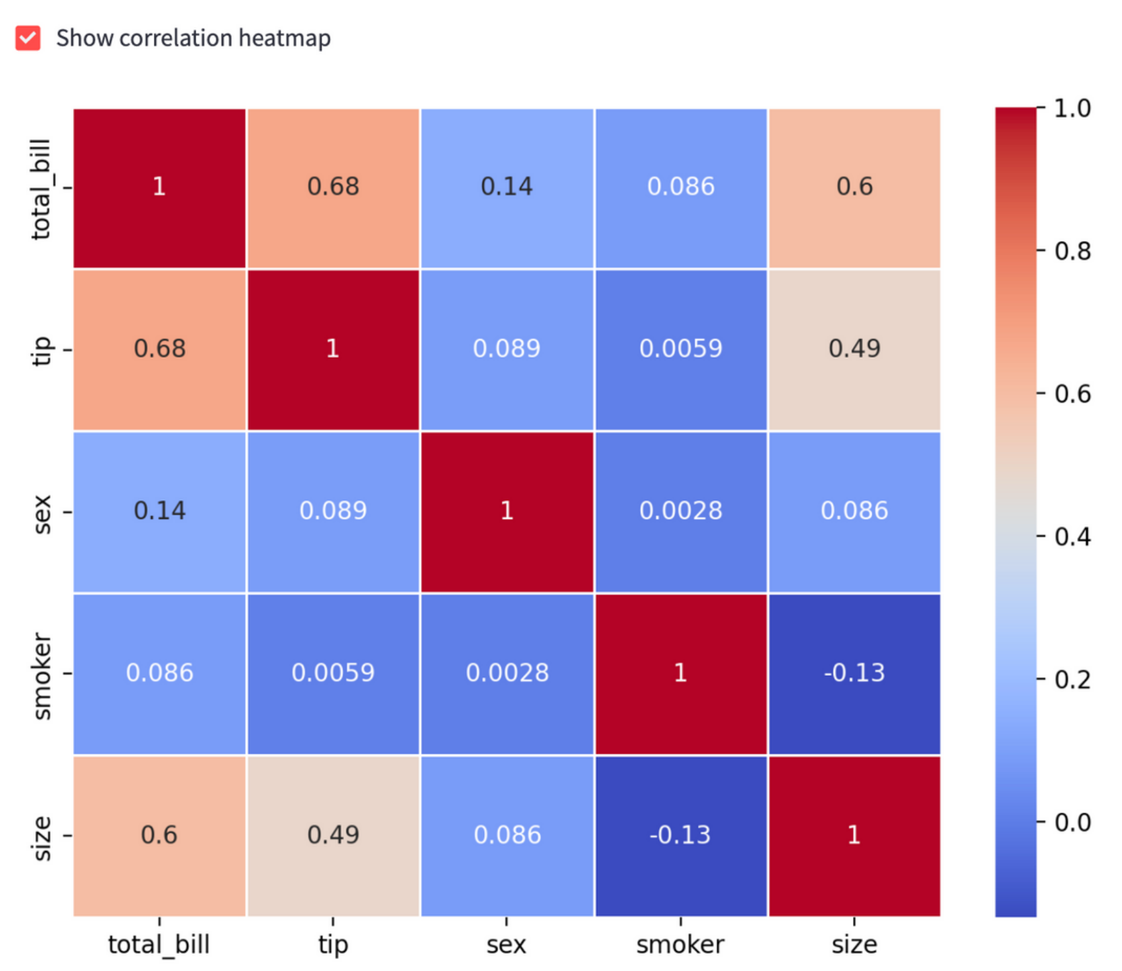
Show summary statistics

	total_bill	tip	sex	smoker	day	time	size
count	244	244	244	244	244	244	244
mean	19.7859	2.9983	0.6434	0.3811	0	0	2.5697
std	8.9024	1.3836	0.48	0.4867	0	0	0.9511
min	3.07	1	0	0	0	0	1
25%	13.3475	2	0	0	0	0	2
50%	17.795	2.9	1	0	0	0	2
75%	24.1275	3.5625	1	1	0	0	3
max	50.81	10	1	1	0	0	6

Q1: WHAT IS THE PERCENTAGE OF CUSTOMERS WHO SMOKE?:

38.11% of the customers in our dataset are smokers.

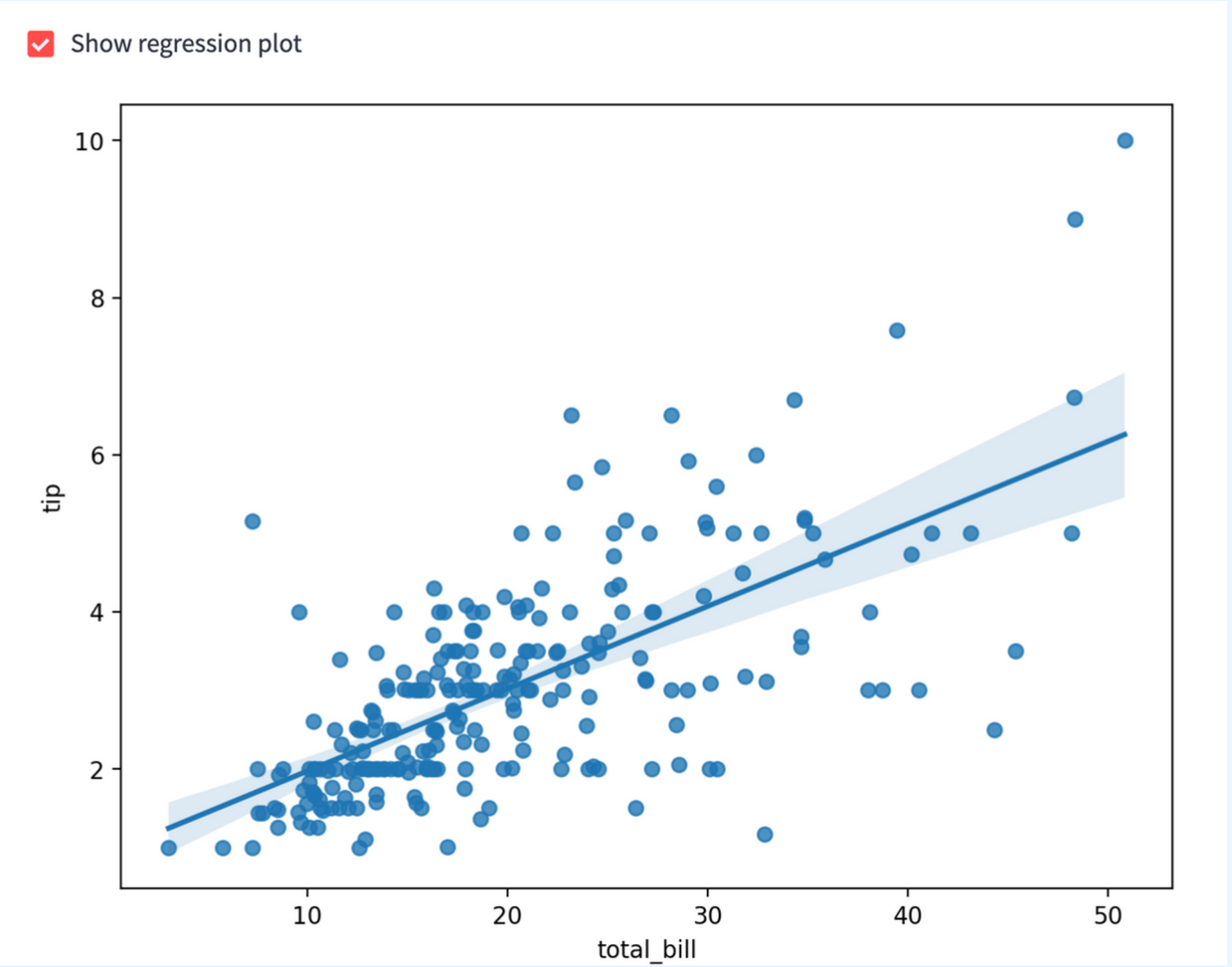
EDA (Exploratory Data Analysis) correlation heatmap



Q2: Correlation Heatmap: How are the factors related to one another?:

The correlation heatmap is used to assess the relationships between the variables. The heatmap revealed that the total bill and tip amount share a strong positive correlation, while the other variables show weaker correlations.

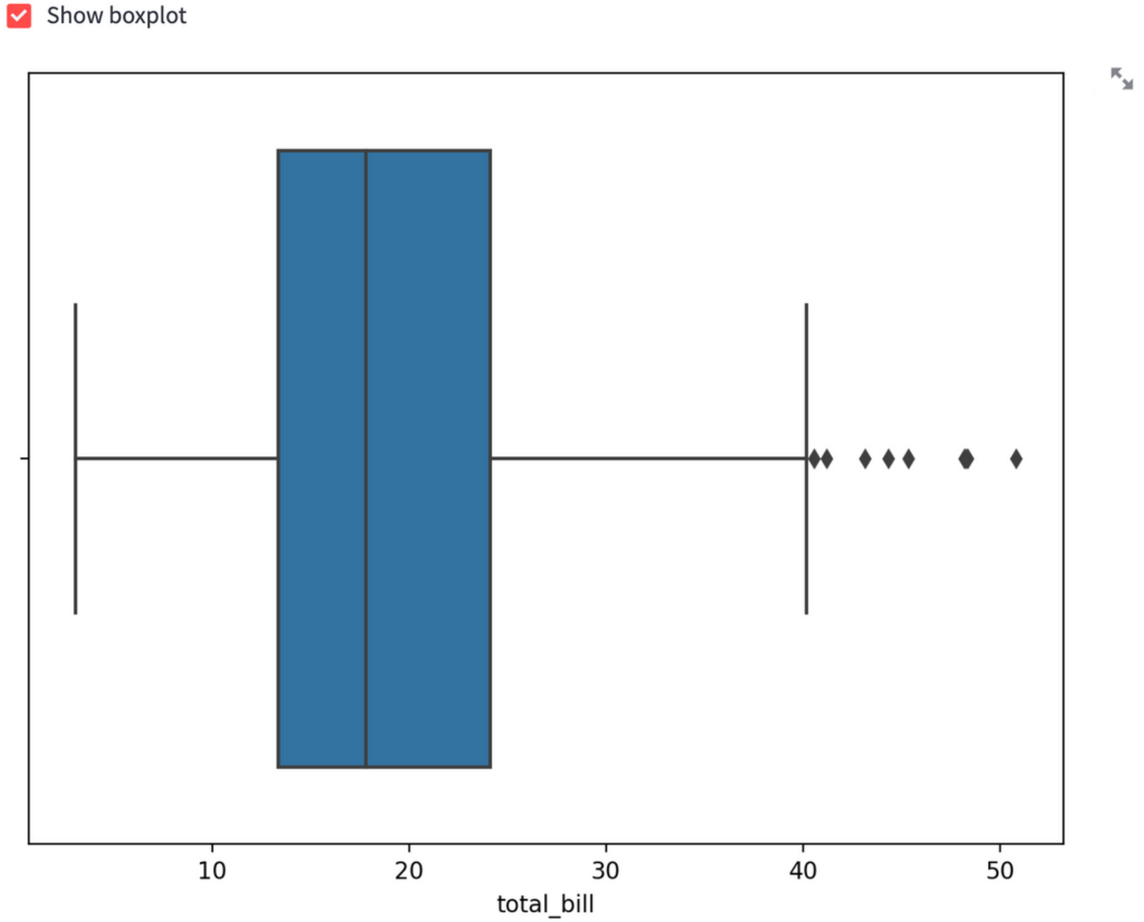
EDA (Exploratory Data Analysis) regression plot



Q3: What is the relationship between the total bill amount and the tip amount as shown in the regression plot?

There is a positive correlation between the total bill amount and the tip amount. This suggests that as the total bill amount increases, the tip amount also tends to increase. This makes sense intuitively, as larger bills usually result in larger tip amounts.

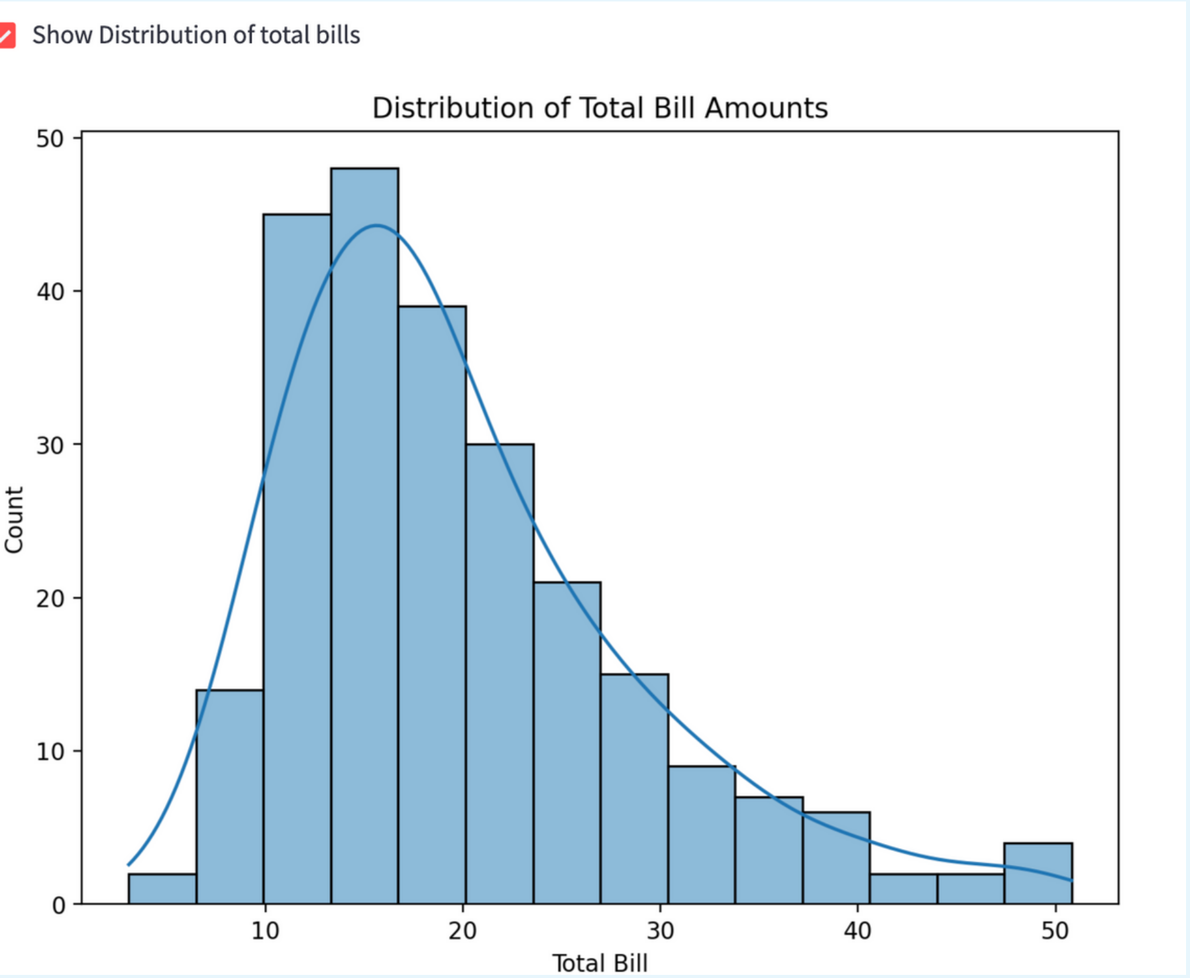
EDA (Exploratory Data Analysis)
Boxplot of total bills



Q4: ARE THERE ANY OUTLIERS IN THE DATA FOR TOTAL BILL?

Using a box plot, there was several identified outliers in the data for total bill amounts, particularly for higher bill values. these outliers removed before proceeding to the predictive modeling phase.

EDA (Exploratory Data Analysis)
Distribution of total bills histogram

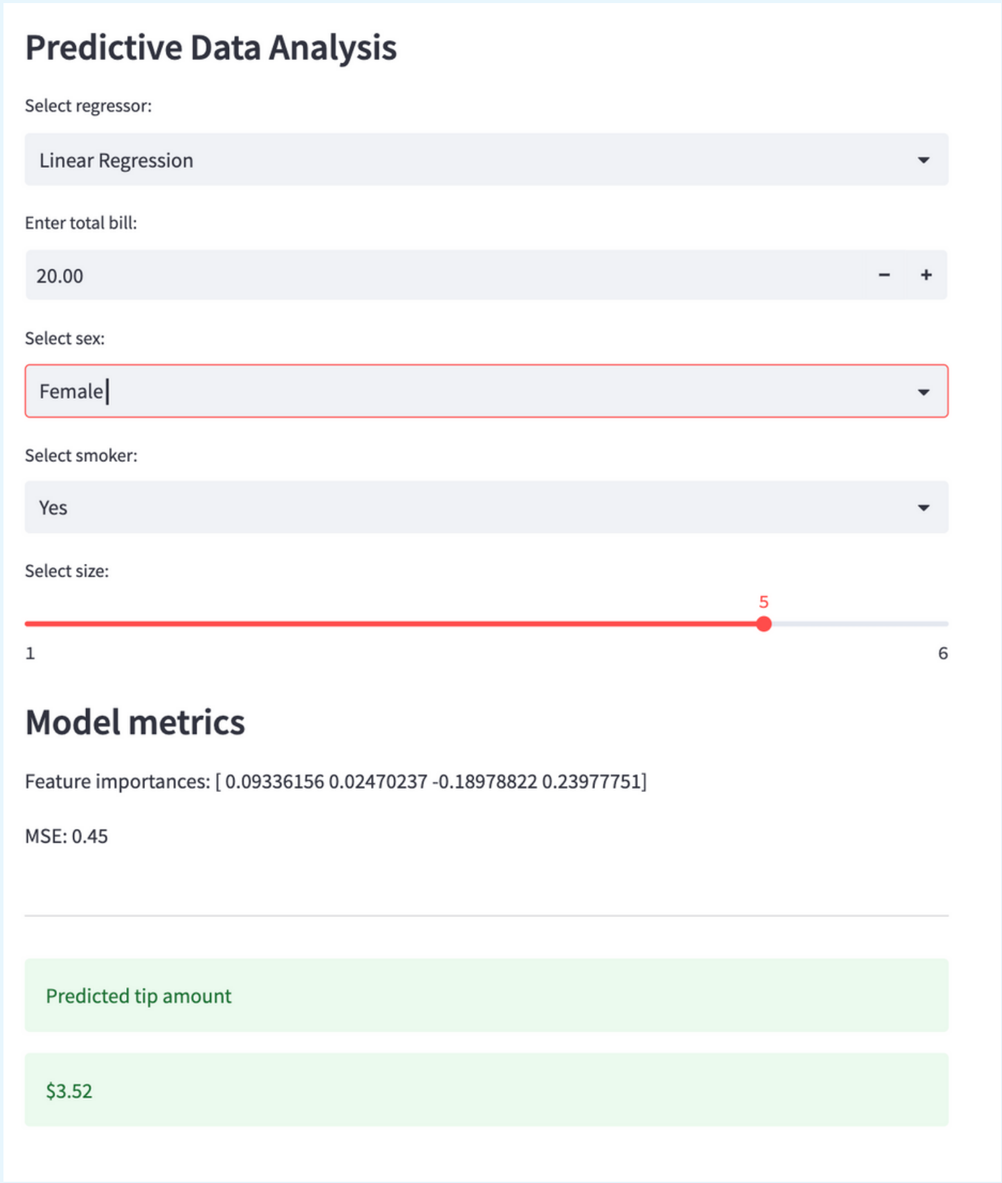


Q5: WHAT IS THE DISTRIBUTION OF TOTAL BILL AMOUNTS?

The histogram visualizes the distribution of total bill amounts. Most bills fell within the $10 to $20 range, with a peak around $15.

PDA (Predictive Data Analysis) Outcomes

- I used the linear regression and random forest regression to predict the waiter's tips depending on the factors: total_bill, sex smoker, size.
- I removed the outliers before making prediction which decreased the MSE(mean squared error in both the case).
- The mean square for Linear is 0.54 and for random forest is 0.72 approx.



Implementation and Deployment(Streamlit)

I used streamlit to deploy and implement my findings.

I used the jupyter notebook for the EDA and PDA and with debugging and rewriting

I built a streamlit app to display the EDA and PDA for the data.