BIG_MART.

MOHAMMED ADHEEB P

MES20MCA-2025

PRODUCT OWNER : PRIYA JD

TABLE OF CONTENTS

INTRODUCTION

- E-commerce or electronic commerce is a process of Buying, selling, transferring, or exchanging products, services, and/or information via electronic networks and computers. Commerce is a fundamental to the success of any business. To streamline trading operations and maintain profits, the industry must focus on commerce, which deals with a lot more than just buying and selling. As the world is digitalizing, ecommerce solutions are increasingly becoming common. The advent of machine learning and artificial intelligence has further enhanced the effectiveness of ecommerce. The examples of ecommerce include eBay, Amazon, Upwork, Olx, etc.
- The advantages of ecommerce shopping are Lower price for products compared to traditional shopping, Time saving, we can avoid crowds in shopping, Available wide range and verity of products, shipping/delivery options and also it provides Feedback option from customers.

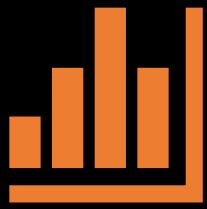
MODULES

- **Steps of this Project:**

1. Business Understanding
    - Description
    - Objective
2. Data Understanding
    - Import Libraries
    - Load Data
    - Statistical summaries and visualizations
        - Mean, standard deviation
        - Heatmap, Distplot, Boxplot, Histplot, Scatterplot, Pairplot, Regplot
3. Data Preparation
    - Missing values imputation
    - Removing Irrelevant Column

4. Exploratory Data Analysis
   - Features Exploration:
   - Distplot, Boxplot, Histplot, Pairplot, Regplot, Barplot

5.Data Splitting: Train test split



6. Log Transformation:

   The log transformation is, arguably, the most popular among the different types of transformations used to transform skewed data to approximately conform to normality. If the original data follows a log-normal distribution or approximately so, then the log-transformed data follows a normal or near normal distribution.

- 7 Modeling:

  - Linear Regression (model 1)

    A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables.

  - Lasso (model 2)

    The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

  - Ridge (model 3)

    Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

- **ElasticNetCv (model 4)**

ElasticNetCV is a cross-validation class that can search multiple alpha values and applies the best one. We'll define the model with alphas value and fit it with xtrain and ytrain data. We can predict xtest data and check the accuracy metrics.

Finding the Prediction Error by MAE (mean absolute error) & RMSE (Root Mean Square Error)

To select the best model.

8. Finalizing Model Choice:

Model performance

Feature importance

9. Deployment:

Predict the annual income of a new record by taking new values to the variables

10. Conclusion & Recommendation

METHODOLOGY

- **Problem Definition:**

A project with an Ecommerce company based in London sells clothing online but they also have in-store style and clothing advice sessions. Customers come in to the store, have sessions/meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want. The company is trying to decide whether to focus their efforts on their mobile app experience or their website. They've asked to help them figure it out.

I worked with the Ecommerce company and the dataset name is Ecommerce.csv. It has Customer info, such as Customer ID, Avg Session length, Time on App, Time on Website, Length of Membership, Yealy amount spent.

Attribute Information:

- Avg. Session Length: Average session of in-store style advice sessions.
- Time on App: Average time spent on App in minutes
- Time on Website: Average time spent on Website in minutes
- Length of Membership: How many years the customer has been a member.
- Yearly Amount Spent: The total amount the customer is spending.

- **Business problem:**

  Interpret which variables are contributing towards the more annual income prediction

- Firstly we use Exploratory Data Analysis (EDA), it is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.
- The next step is Data Preparation which includes : Data cleaning, Feature engineering, Data transformation
- Then I build 4 models from it, and find out the errors after some aggregations are applied to these models (log transformation)
- Finally get the best model out of these models with low mae & mape
- Find out which variables are contributing towards the more annual income prediction.
- Predict the annual income of a new record by taking new values to the variables

Sample Data

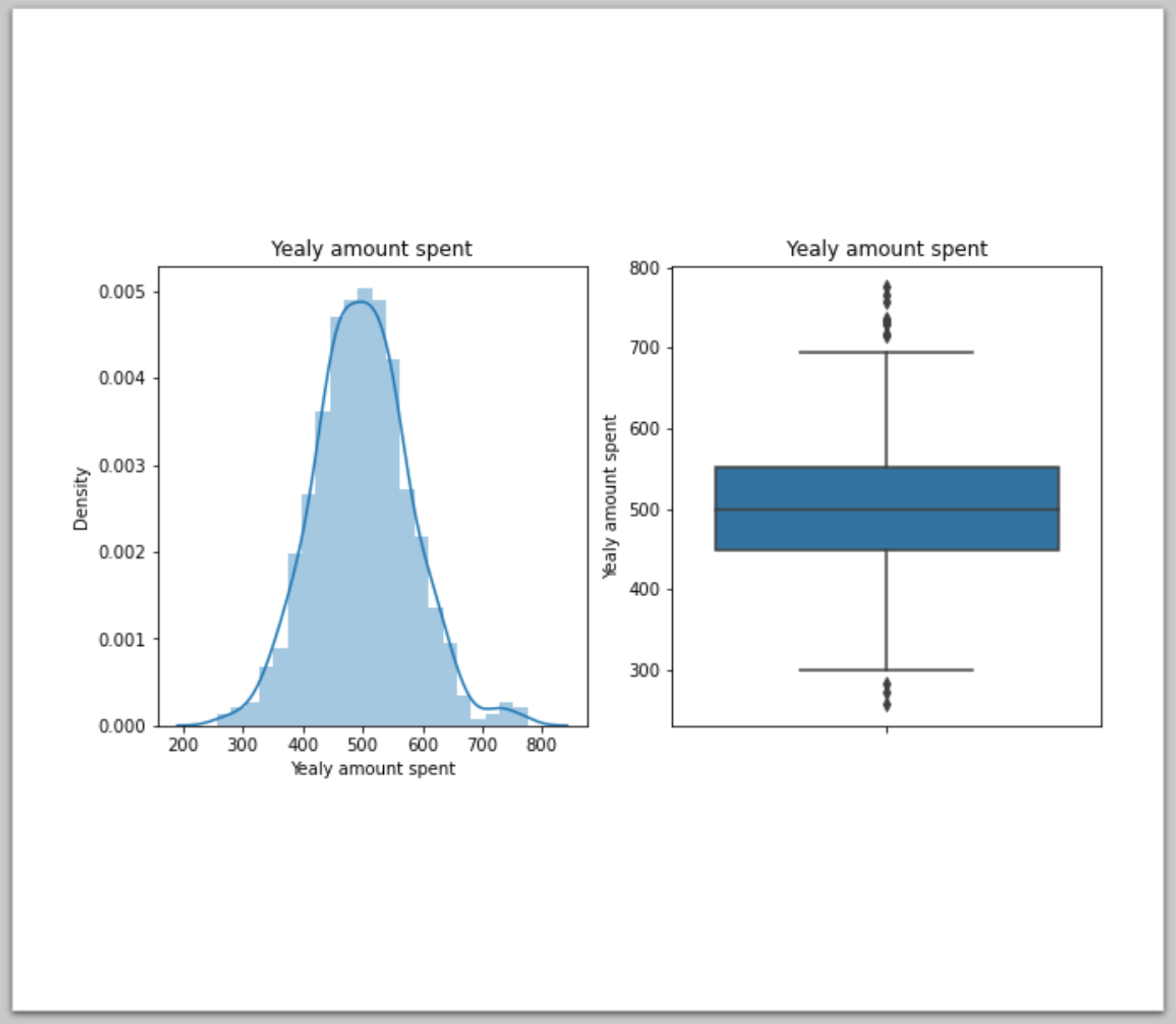
	Customer ID	Avg Session length	Time on App	Time on Website	Length of MemberShip	Yealy amount spent
0	1	32.538350	12.028846	35.850540	3.648854	576.098273
1	2	34.206718	12.226851	37.246443	1.987536	460.784955
2	3	31.535240	11.814341	36.610697	3.351191	349.739791
3	4	32.199577	11.295163	37.425695	5.234337	547.709921
4	5	33.570137	13.500972	36.856165	3.938603	487.055641

Statistical summaries

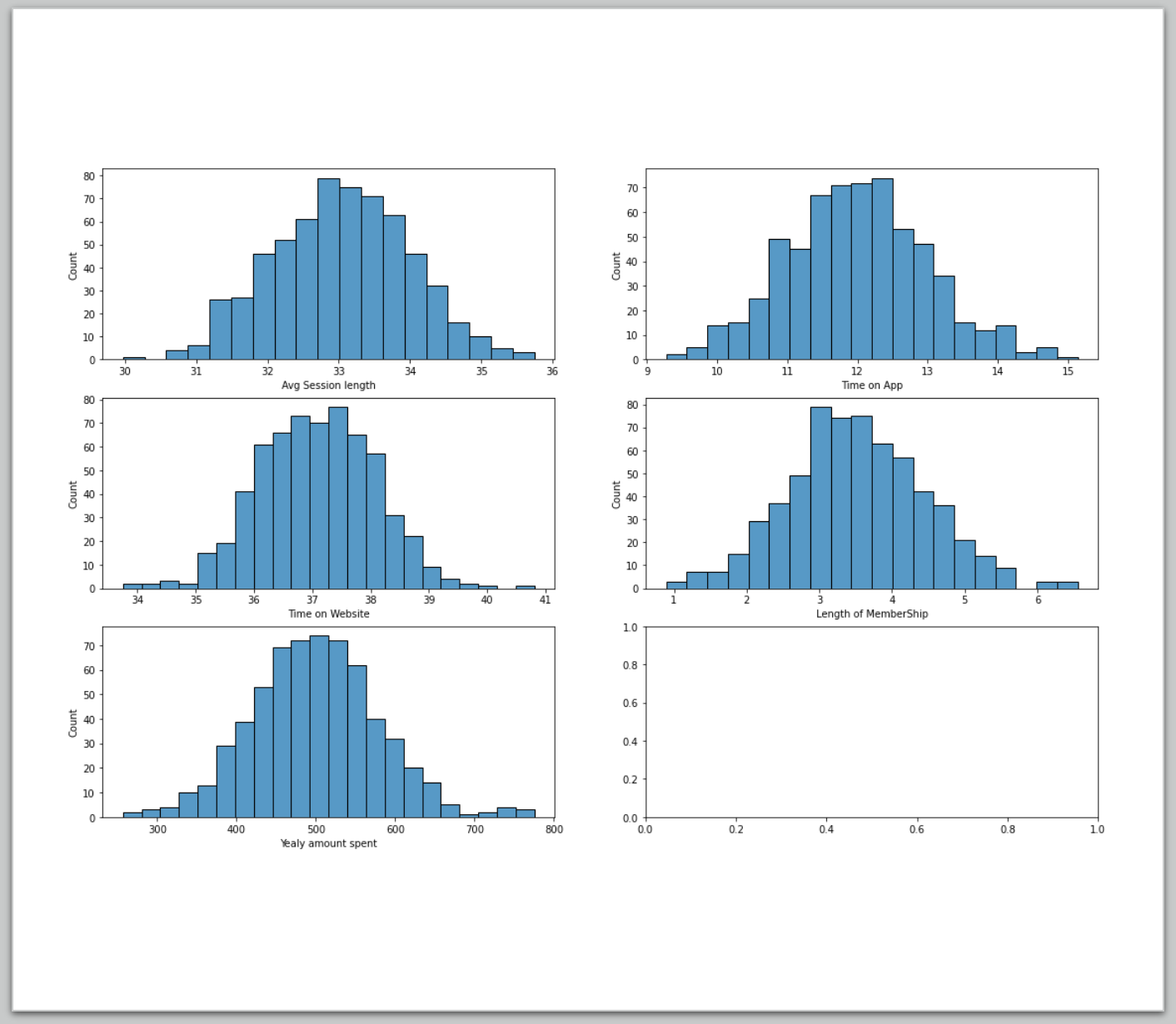
	count	mean	std	min	25%	50%	75%	max
Customer ID	623.0	312.000000	179.988889	1.000000	156.500000	312.000000	467.500000	623.000000
Avg Session length	623.0	33.039429	0.949071	29.972300	32.375680	33.044399	33.700947	35.744415
Time on App	623.0	12.001459	0.996609	9.273248	11.334163	11.998823	12.647695	15.138317
Time on Website	623.0	37.079018	0.991753	33.751071	36.376725	37.113631	37.773880	40.808388
Length of MemberShip	623.0	3.514850	0.948063	0.891398	2.906558	3.504771	4.140305	6.553916
Yealy amount spent	623.0	499.600023	80.032965	256.670000	447.665160	498.806136	551.257208	775.337626

Exploratory Data Analysis

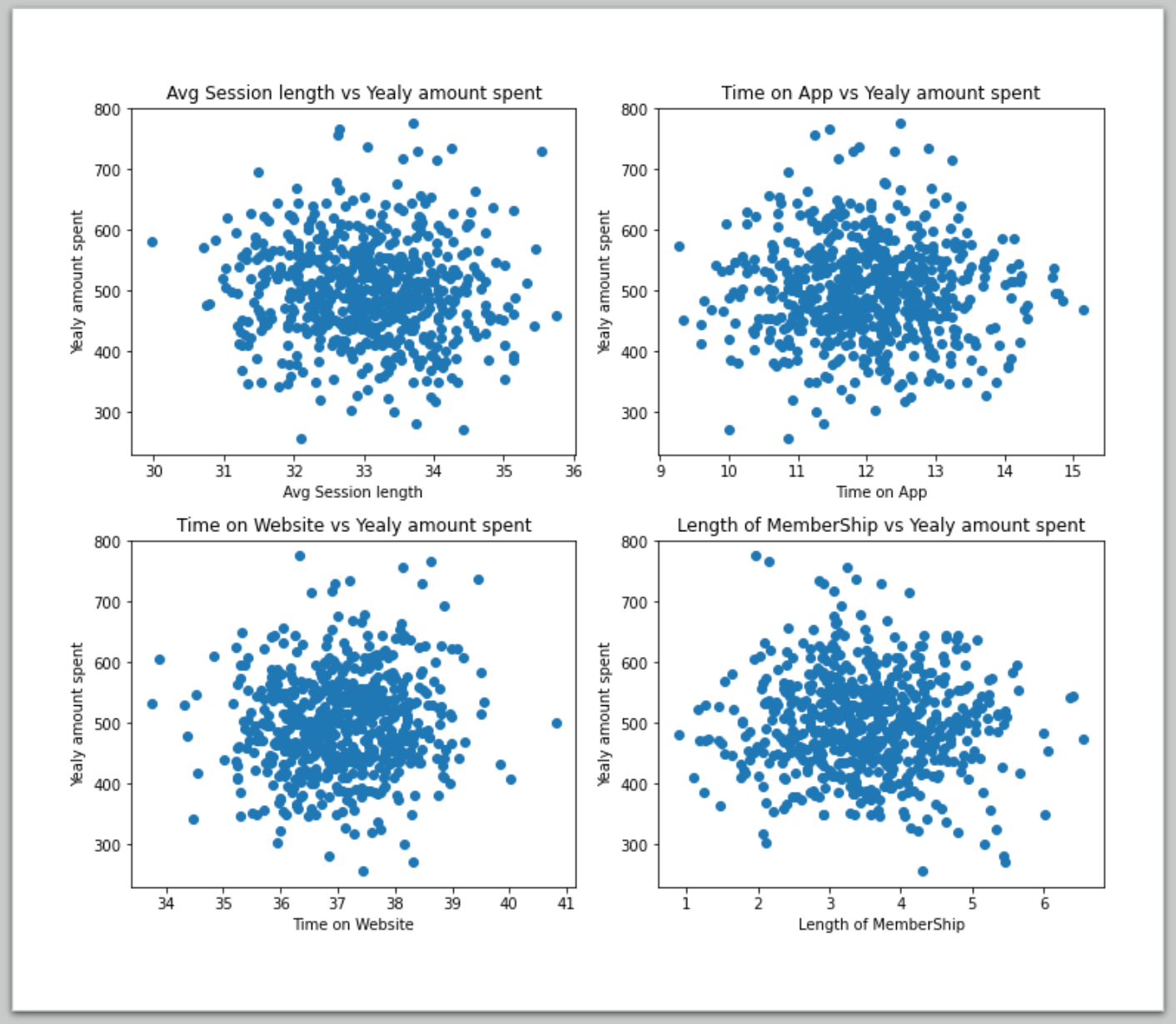
- Boxplot
- Distplot



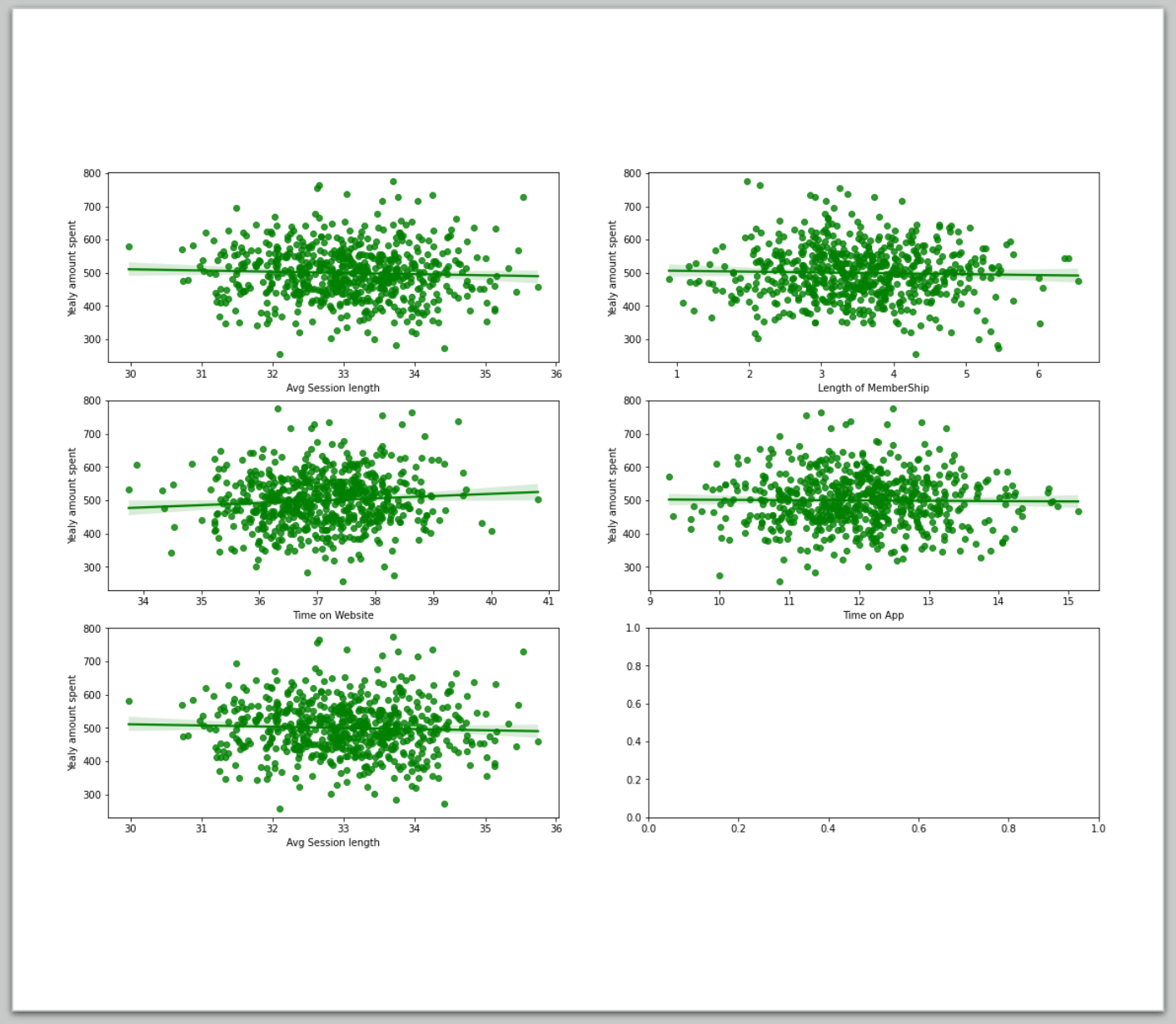
- Histogram


Count
Avg Session length
Time on App
Time on Website
Length of MemberShip
Yealy amount spent

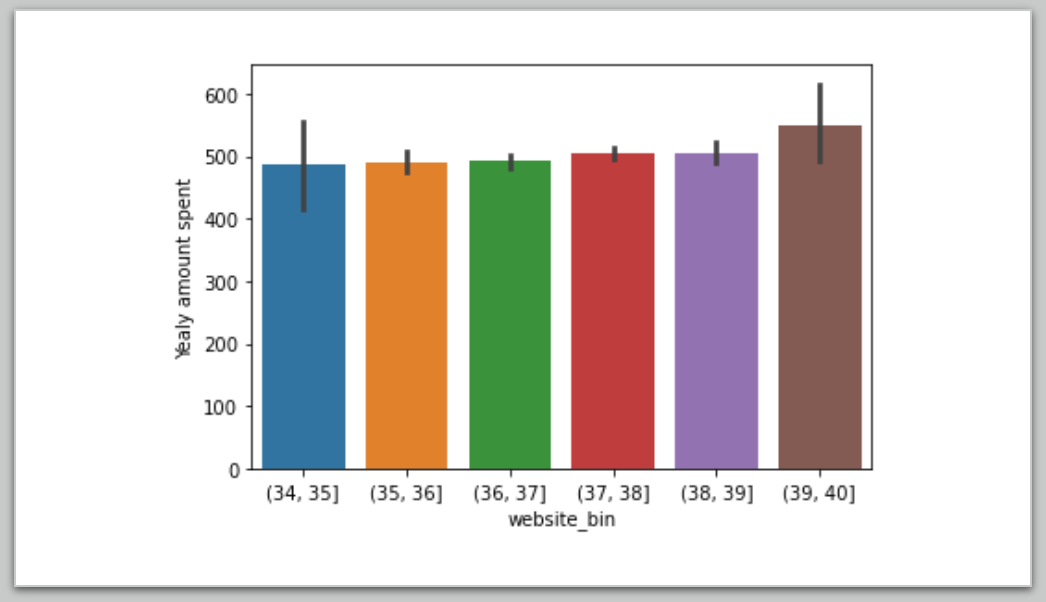
- Scatterplot

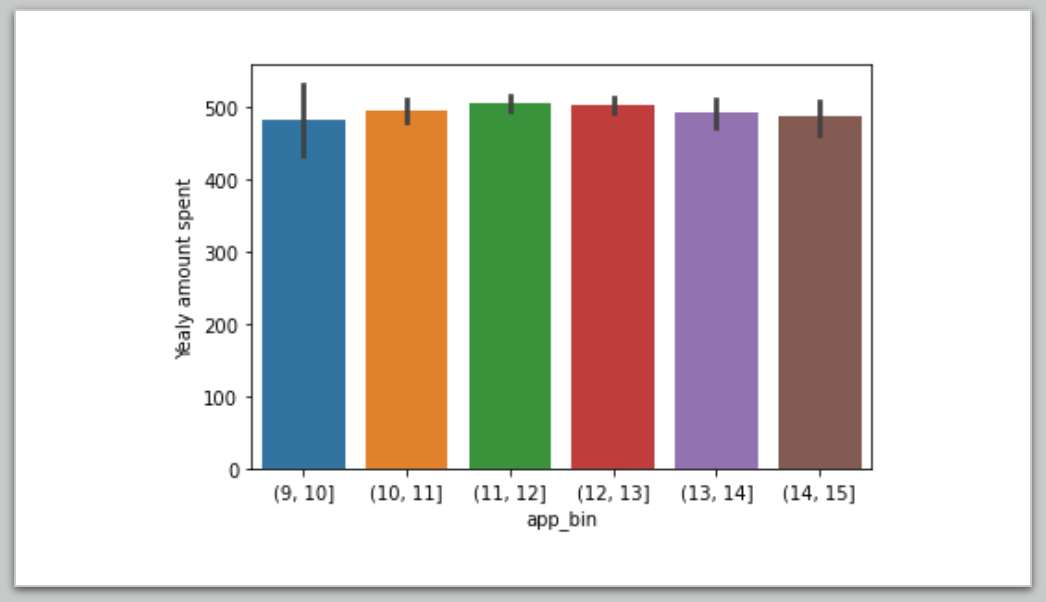


- Regplot

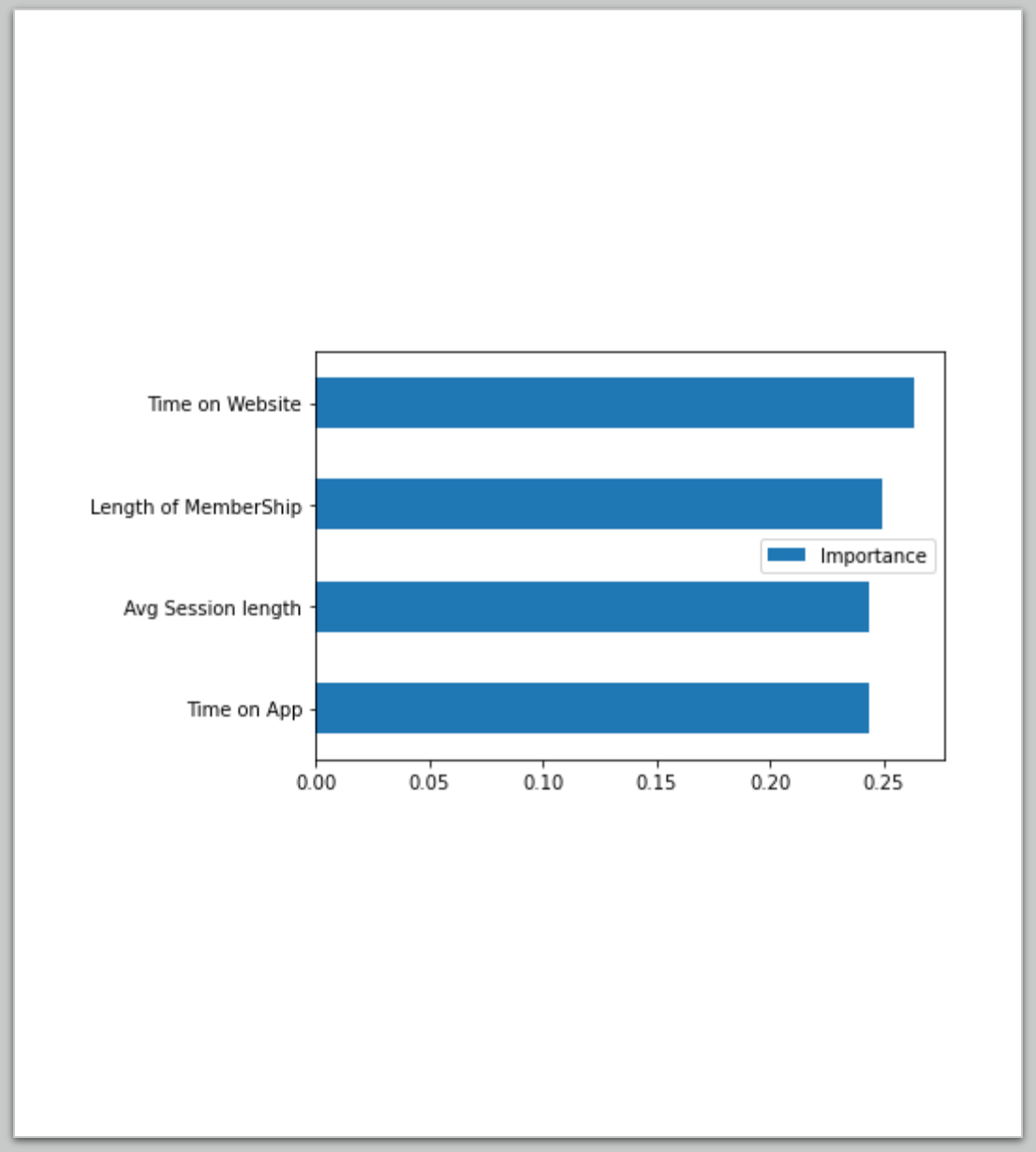

Yealy amount spent
Avg Session length
Length of MemberShip
Time on Website
Time on App
Avg Session length

- Barplot


600
500
400
300
200
100
0
Yealy amount spent
(34, 35]
(35, 36]
(36, 37]
(37, 38]
(38, 39]
(39, 40]
website_bin


500
400
300
200
100
0
Yealy amount spent
(9, 10]
(10, 11]
(11, 12]
(12, 13]
(13, 14]
(14, 15]
app_bin

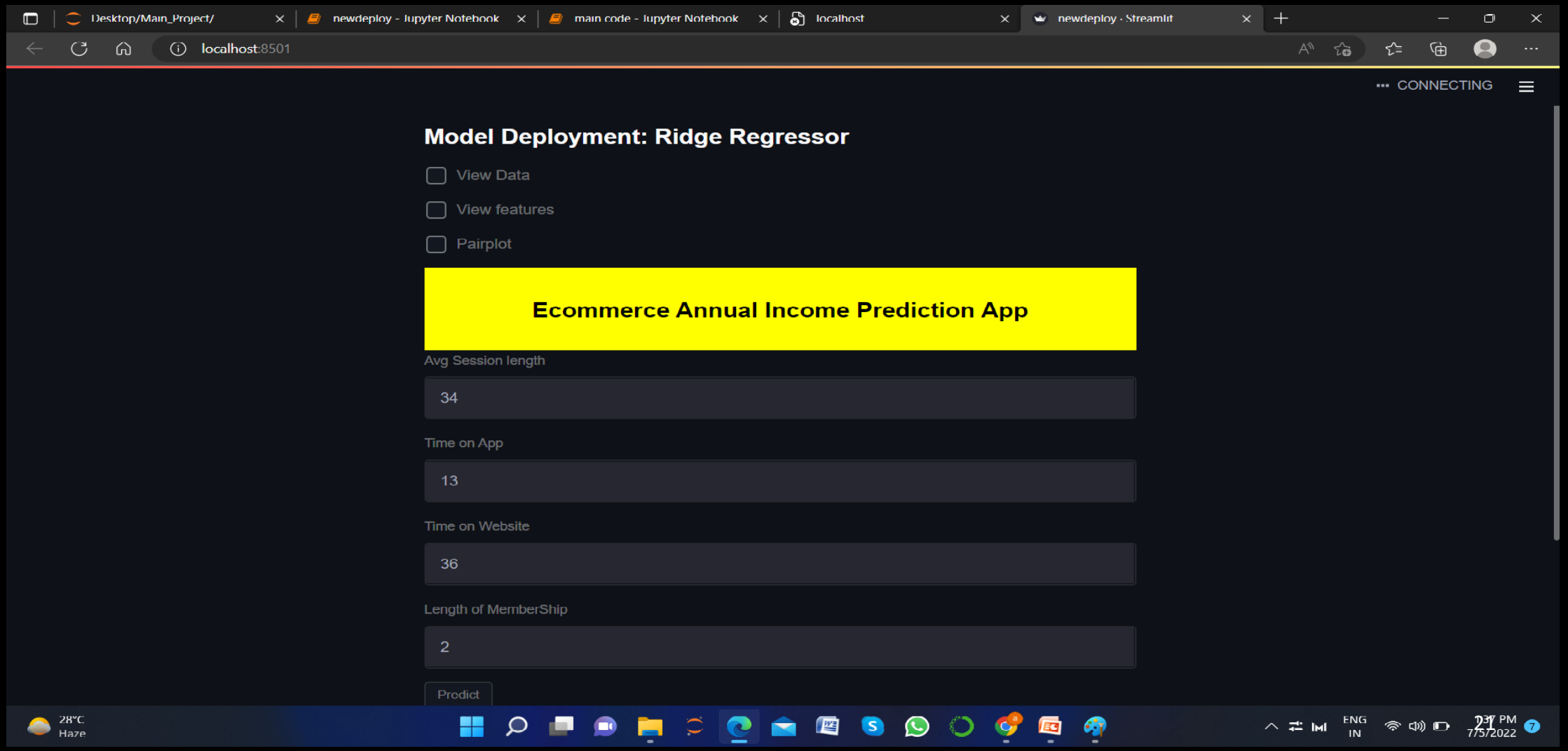
- **Feature importance**

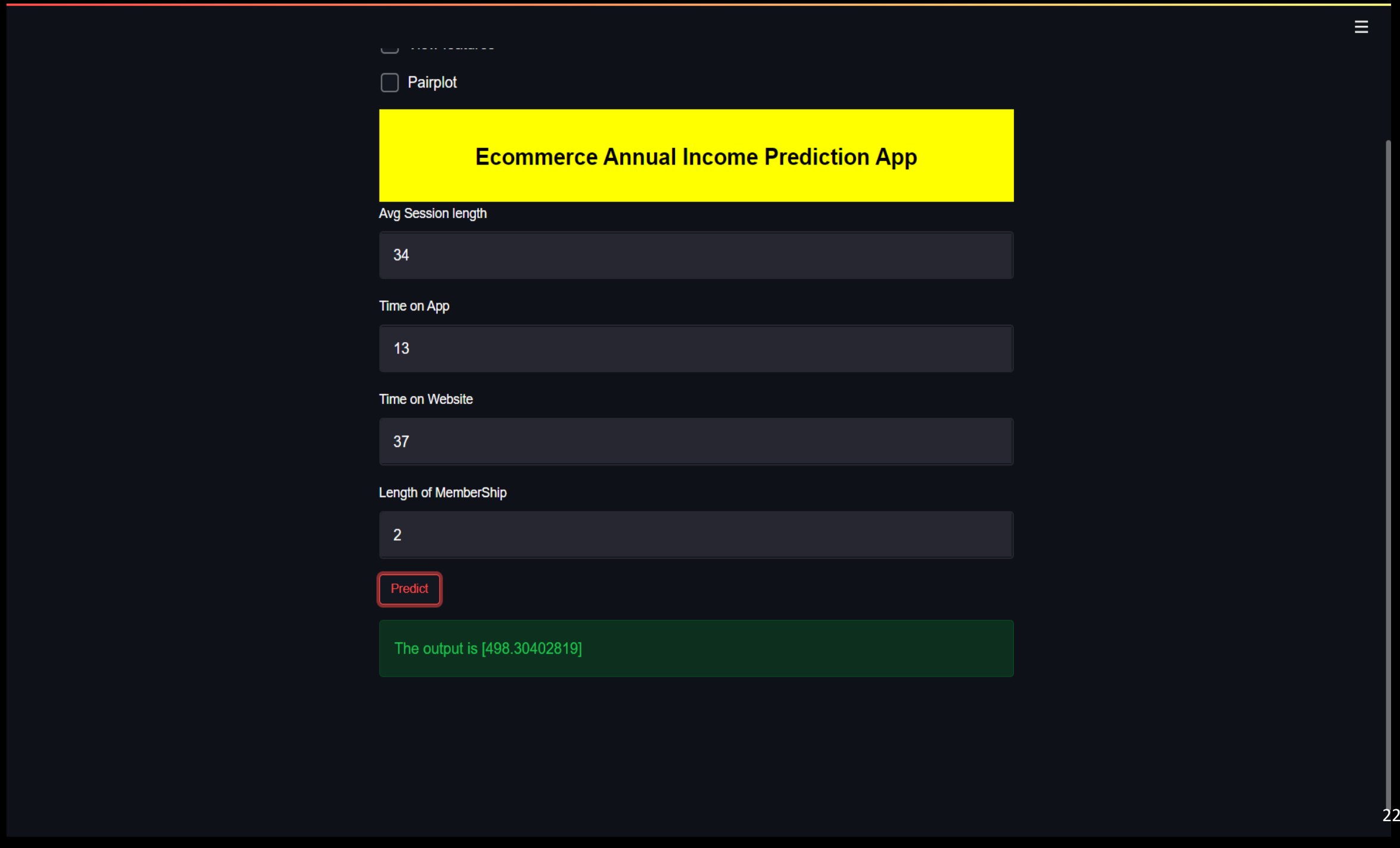


- MAE & RMSE ERRORS

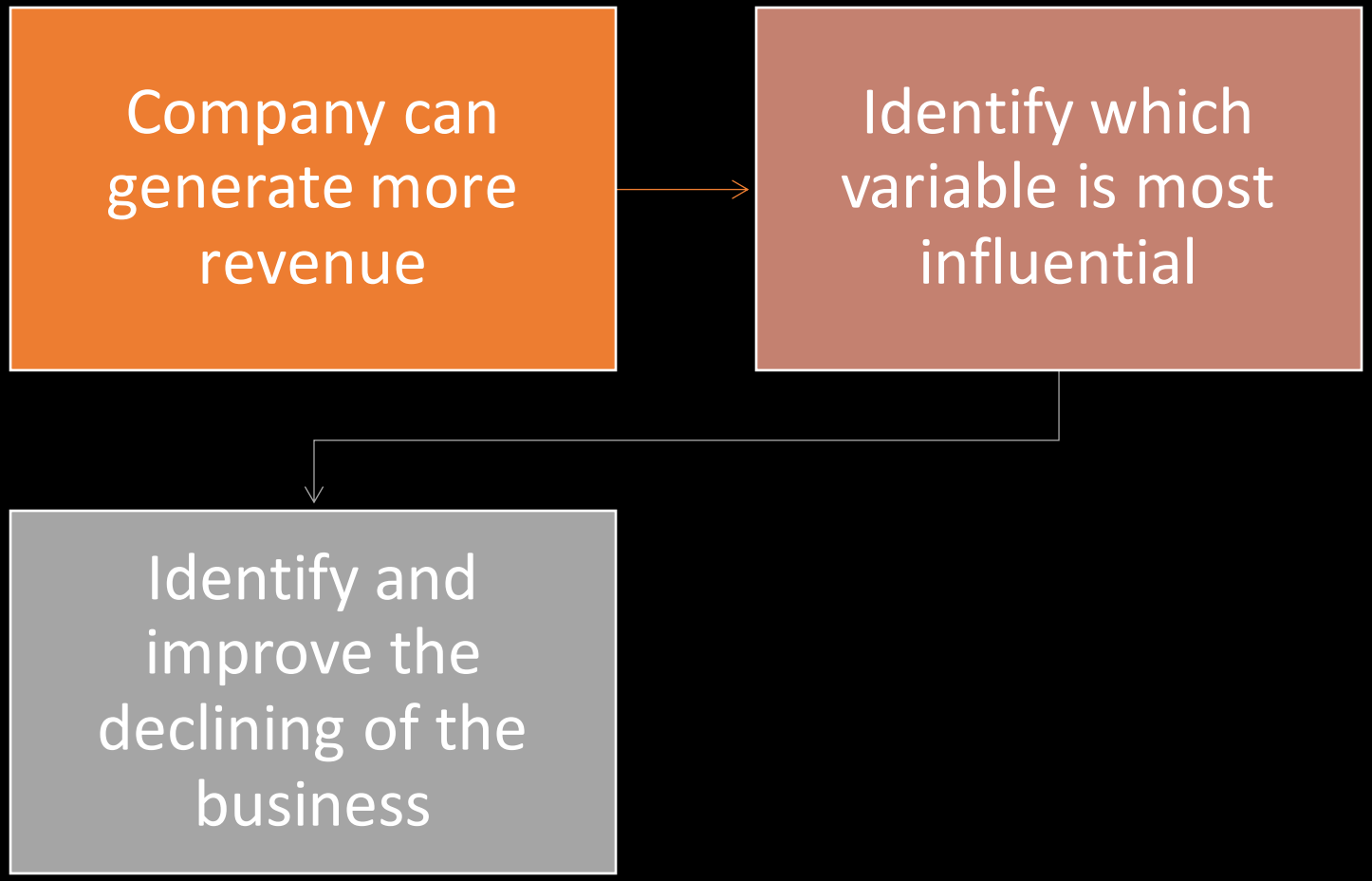
	MAE_test	MAE_train	RMSE_test	RMSE_train
ElasticNetCV	60.012287	62.918013	73.447169	80.585799
Lasso	60.056694	62.884589	73.460515	80.592627
Ridge	60.382623	62.892897	73.847466	80.565753
LinearRegression	60.387280	62.892734	73.852772	80.565750

Deployment


Desktop/Main_Project/
newdeploy - Jupyter Notebook
main code - Jupyter Notebook
localhost
newdeploy - Streamlit
localhost:8501
CONNECTING
Model Deployment: Ridge Regressor
View Data
View features
Pairplot
Ecommerce Annual Income Prediction App
Avg Session length
34
Time on App
13
Time on Website
36
Length of MemberShip
2
Predict
28°C
Haze
ENG
IN
7/7/2022


Pairplot
Ecommerce Annual Income Prediction App
Avg Session length
34
Time on App
13
Time on Website
37
Length of MemberShip
2
Predict
The output is [498.30402819]

FUTURE ENHANCEMENT



DEVELOPING ENVIRONMENT

Hardware Requirements

- Processor - Intel Core i5 (min)
- Speed - 1.5 GHz (min)
- RAM - 8 GB (min)
- Hard Disk - 100 GB or (min)
- GPU - 1 GB (min)

Software Requirements

- Language :Python
- Front end : Python Jupyter
- Operating system : windows 8 or above
- IDE : Anaconda Navigator
- Dataset: 2020 sales result

PROJECT PLAN

User Story ID	Task Name	Start Date	End Date	Hours	Status
1	Sprint 1	20/04/2022	2/05/2022	19	Completed
2		4/05/2022	14/05/2022		Completed
4	Sprint 2	16/05/2022	20/05/2022	10	Completed
5		22/05/2022	24/05/2022		Completed
6	Sprint 3	25/05/2022	5/06/2022	7	Completed
8	Sprint 4	6/06/2022			In progress

USER STORY

User Story ID	As a type of User	I want to <perform some task>	So that I can < Achieve Some Goal>
1	ADMIN	Predict the Annual Income of a new record	View the annual income of new record

PRODUCT BACKLOG

Sl.No	Priority <High / Medium / Low>	Size (Hours)	Sprint <#>	Status <Planned / In progress /Completed>	Release Date	Release Goal
1	High	6	1	Completed		Data Understanding
2	High	6		Completed		Data Preparation
3	High	4		Completed		EDA on Distplot & Boxplot
4	Medium	6	2	completed		Feature Exploration on Pairplot
5	High	2		Completed		Feature Exploration on Regplot
6	High	2		completed		Data Splitting
7	High	10	3	Completed		Modeling
8	Low	2	4	Incomplete		Web Page Design
9	High	6		Incomplete		Final Prediction
10	High	1		Incomplete		Raw file conversion

SPRINT PLAN

Backlog Item	Status And Completion Date	Original Estimation in Hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
			hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs	hrs
Data Understanding	2/05/2022	6	1	1	1	1	0	1	0	0	1	0	0	0	0	0
Data Preparation	8/05/2022	6	1	1	0	1	1	0	1	1	0	0	0	0	0	0
Exploratory Data Analysis	14/05/2022	4	1	0	0	1	1	0	0	0	0	0	0	0	0	0
Feature Exploration	20/05/2022	8	0	2	1	0	2	0	3	0	0	0	0	0	0	0
Data Splitting	24/05/2022	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Modelling	01/06/2022	6	3	3	0	0	0	0	0	0	0	0	0	0	0	0
Full Training	04/06/2022	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
Deployment(Prediction)	25/06/2022	5	3	2	0	0	0	0	0	0	0	0	0	0	0	0
Total		41	11	8	2	3	4	1	4	1	1	0	0	0	0	0

SPRINT ACTUAL

Backlog Item	Status And Completion Date	Original Estimation in Hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
			hrs	hrs	hrs	hrs	hrs	hrs
Data Understanding	2/05/2022	6	0	1	1	1	0	3
Data Preparation	8/05/2022	7	1	0	0	0	3	3
Exploratory Data Analysis	14/05/2022	5	1	0	1	1	1	0
Feature Exploration	20/05/2022	10	1	1	1	1	2	4
Data Splitting	24/05/2022	3	1	1	0	1	0	0
Modelling	01/06/2022	9	3	3	3	0	0	0
Full Training	04/06/2022	1	1	0	0	0	0	0
Deployment(Prediction)	30/07/2022	6	2	2	2	0	0	0
Total		47	8	6	6	4	6	10

THANK YOU