CSE3505 - Foundations of Data Analytics

Class ID - CH2022231000351

Slot - L37+L38

Type of the Project - *C*

J Component – Review Project Report

Fast Food Demand Analytics and Prediction

By

19MIA1005	Mohit More
19MIA1066	Madasu Deepika
19MIA1069	G. Harinisri

M.Tech CSE with Specialization in Business Analytics

Submitted to

Dr. Priyadarshini,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

DECLARATION

We hereby declare that the report titled "***Fast Food Demand Analytics and Prediction***" submitted by us to VIT Chennai is a record of bonafide work undertaken by us under the supervision of Dr. Priyadarshini, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai.

Signature of the Candidate
Mohit more (19MIA1005)
Madasu Deepika(19MIA1066)
G.Harinisri(19MIA1069)

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr.R.Priyadarshini, School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work. We are extremely grateful to Dr. R. Ganesan, Dean, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for giving all of us the exposure to this course and enhancing the knowledge.

We express thanks to our Head of the Department for his support throughout the course of this project.We also take this opportunity to
thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

Abstract

In this paper, demand forecasting in restaurants using machine learning is proposed. Many researches have been proposed on demand forecasting technology using POS data. However, in order to make demand forecasts at a real store, it is necessary to establish a store-specific demand forecasting model in consideration of various factors such as the store location, the weather, events, etc. Therefore, we constructed a demand forecasting model that functionally combines the above mentioned data using machine learning. In this paper, the demand forecasting model using machine learning and the verification result of the model using real store data is discussed. In this research is used to evaluate the factors of customer satisfaction highlighting the fast-food trend. Furthermore, this research will also allow the reader to know which factor leads to customer satisfaction and also to determine the various factors which influence visiting fast-food outlets. In addition, the research sought to discover how satisfied customers decide the trend of fast-food restaurants, what factors make customers satisfied and why they prefer a specific fast-food outlet.

Introduction

A fast-food restaurant is a growing business and with time evolving the pattern of eating habits is changing. Therefore, many people have started opting for fast food which is an appealing phenomenon for them.
The prominent reasons contributing to the growth of the Fast the food industry includes the aspects like increasing disposable income, also more options available in the cuisines and consumer's interest in experimenting with all possible combinations of Fast food.
The concept of fast-food is prevailing since quite a few years now. In this fast-moving world where the expenses are day by day inclining, it is a must for all the members of the family to work for living. Also many times lot of people shift to other region away from the family in order to earn a living. This kind of a rush lifestyle gives a scope for consumption of quick meals which are easily prepared and quickly served which we term as fast-food. This fast-food at the time of being prepared in no time it is also a pleasure to the tongue. Therefore, fast-food industry is widening day by day which is leading to its progress along with certain drawbacks. Food served in fast food restaurants typically caters to a "meat-sweet diet" and is offered from a limited menu; is cooked in bulk in advance and kept hot which is finished and packaged to order and is usually available ready to take away, though seating may be provided.

Claims

The core agenda of this project is to evaluate the factors of customer satisfaction highlighting the fast-food trend. Furthermore, this project will allow us to know which factor leads to customer satisfaction and also to determine the various factors which influence visiting fast-food outlets. In addition, to this we can discover how satisfied customers decide the trend of fast-food restaurants, what factors make customers satisfied and why they prefer a specific fast-food outlet.

To create an app which will predict the fast food prices and the data analysis of overall consumption.

Literature Review

Sno	TITLE	AUTHOR & YEAR OF PUBLICATION	FINDINGS
1	Demand forecasting in restaurants using machine learning and statistical analysis	TakashiTanizaki TomohiroHoshino TakeshiShimmura TakeshiTakenaka 2019	Demand forecasting in restaurants using machine learning was proposed. Forecasting rate for boosted decision tree is low and other algorithms did not have much difference. The forecast rate did not exceed 85%.
2	Food Demand Prediction Using the Nonlinear Autoregressive Exogenous Neural Network	Krzysztof Lutoslawski Marcin Hernes Joanna Radomska Monika Hajdas Ewa Walaszczyk Agata Kozina 2021	Data science methods, including artificial intelligence methods, was used. The aim of this research was to develop models for food demand prediction based on a nonlinear autoregressive exogenous neural network. The architectures of the developed models differed in the number of hidden layers and the number of neurons in the hidden layers, as well as with different sizes of the delay line, were tested for a given product.
3	Predicting food demand in food courts by decision tree approaches	Ahmet SelmanBozkir Ebru AkcapinarSezer 2010	Three decision tree methods CART, CHAID and Microsoft Decision Trees are utilized. As a result, prediction accuracies up to 0.83 in R2 are achieved.

4	Data mining on time series: an illustration using fast-food restaurant franchise data	Lon-MuLiu SiddharthaBhattacharyya L.Sclove RongChen William J.Lattyak 2001	This shows how data mining can be applied to such time series, and help the franchise reap the benefits of such an effort. Time series data mining at both the store level and corporate level are discussed. Outlier detection also leads to information that can be used not only for better inventory management and planning, but also to identify potential sales opportunities as a part of results.
5	Predicting consumer preference for fast-food franchises: a data mining approach	Y Hayashi, M-H Hsieh R Setiono 2009	They evaluated the adequacy of two data mining techniques, decision tree and neural network in analysing consumer preference for a fast-food franchise.

			The generated rules show that while both decision tree and neural network models can achieve predictive accuracy of more than 80% on the training data samples and more that 70% on the cross-validation data samples
6	FOOD DEMAND PREDICTION USING MACHINE LEARNING	K.Aishwarya,Aish war ya.N.Rao, Nikita Kumari, AkshitMishra, Mrs.Rashmi M R 2020	The demands depend upon many explicit and hidden context such as season, region etc. The number of order is used to forecast stock of items, using machine learning with internal and external data. They used an appropriate algorithm for demand. Algorithms like Bayesian Linear Regression, LASSO, XGBoost algorithm are used that considerably improves the forecasting performance.

Methodology

Lack of fast-food fulfillment to the consumer, excesses of fast food over the estimated demand, and business loss profit caused by inaccurate demand prediction are common nowadays in fast food centers and fast food-based businesses (based on local context - Sri Lanka).

Therefore, proposes a solution to avoid this problem by predicting consumer demand for the fast-food sector. Used a forecasting algorithm known as Cat-Boost with a data categorization technique. Fast food demand is affected by several independent variables such as seasonality, trend, price fluctuation, and length of historical data.

A combination of these selected variables was used to calculate demand prediction using parameter tuning in the CatBoost algorithm and other algorithms (slightly different but in the same domain) used for the experiment (Such as Linear Regression, LGBM, and XGBoost). However, CatBoost was the best-performing model that was selected. Therefore, windows-based standalone solution was developed to yield fast-food demand prediction statistics.

Dataset Description:

Dataset derived from the Kaggle platform (https://www.kaggle.com/ghoshsaptarshi/av-genpact-hack-dec2018). One dataset is a combination of three single information files. One file consists of historical demand information for each center, another file consists of data center information and another file consist of meal information. Auxiliary file related to testing information also used for the demand prediction.

1. Historical Demand Information file – "trainForLearnInformation.csv"

This file consists with the historical information of demand for each center. Defined variables listed with a brief description in below,

base_price - Average price of the meal
checkout_price - Sold price of the meal
meal_id - Id of the meal
center_id - Id of the meal center
week - Week number of the sold meal

id - Id of the record
emailer_for_promotion - Removed this variable from the implementation
homepage_featured - Removed this variable from the implementation
num_orders - Demand for the meal

1. Historical center information file – centerInformation.csv

This file consists of historical data for each center. Defined variables listed with a brief description in below,

center_id - Id of the meal center
city_code - Code of the center located city
region_code - Removed this variable from the implantation
center_type - Centre type
op_area - Removed this variable from the implementation

1. Historical meal information file – mealInformation.csv

This file consists historical data of meal information. Defined variables listed with a brief description in below,

meal_id - Specific Id for the meal
category - Categorized name for the meal
cuisine - Type of the cuisine for the meal

4.Test Information file data from 146th week to 155th week – testInformation.csv

This file consists of test data for model validation purpose. Same variables included as mentioned in the historical demand information file (trainForLearnInformation file) except for target variable “num_orders”.

MODULES

1) **Data Preprocessing from the Extracted Data**

When the user inputs files to the system, all information on the submitted files is merged into a separate dataset. Therefore, not having null values or missing values is compulsory. After this, validation process will initiate. Therefore, merging information is required to be validated. As an example, when merging file called “trainForLearnInformation”, specific variable name and their contents should be matched to the same name and contents in the other file. Such as trainToLearn file “meal_id” should match in the mealInfromation file “meal_id” variable.

2) **Exploratory Data Analysis for the Dataset**

This is a process, performing an initial investigation on data. Such as identify patterns, identify anomalies, test hypotheses and validate assumptions with the help of graphical statistics representation or statistical information. At the beginning of the exploratory data analysis process, it is compulsory to identify and remove unnecessary variables that are not contributing to the prediction process. Therefore, four variables were removed. Such as region_code, op_area, emailer_for_promotion and homepage_featured. emailer_for_promotion and homepage_featured column with data were dropped by updating the file. However, region_code, op_area kept in the dataset for identifying its necessity for implementation.The next process of the exploratory data analysis is the standardization of features. Standardization is a technique. It changes the values of numeric columns in the dataset to a common scale, without altering differences in the ranges of values.

3) **Feature Engineering from the Extracted Information**

This is a method to create features according to the domain knowledge that enables to enhancement performance and accuracy of the machine learning models using the dataset.

4) **Data Transformation for Eliminate Outliers**

In the demand prediction context, it is compulsory to outlier data to be 0% on a targeted variable called "num_orders". Therefore, this necessity is achieved by using the Interquartile range method. Log transformation is the most popular among the different types of transformations used to transform skewed data to approximately conform to normality in feature engineering. Therefore, the target variable called "num_orders" is not aligned with normality and non-use of transformation methods will reduce the performance of the data model. Therefore, it was decided to include log transformation on the targeted variable "num_orders".

5) **Machine Learning Algorithms for Demand Prediction**

Multiple data modeled using gradient boosting algorithms like CatBoost algorithm. Those algorithms are implemented with feature extraction, data transformation and data preprocessing for achieving better accuracy on the predicted result.

After the above process, it was decided to categorize dataset "week" values into a created feature called "Quarter" and "Year" as shown in figure 16. Reason for categorizing, train dataset contains 146 weeks of data which is approximately 11 quarters and one-quarter consists of approximately 13 weeks. That is the reason for the week divided by 13 for quarter and purpose of the calculation it was defined to 12 quarters. And year consists of approximately 52 weeks. Therefore, when it comes to years, it was identified 3 years of data. That is the reason for the week divided by 52. The goal of mapping those related data using the map method is to return a list of the results according to the calculated outcome. Then manipulated those data accordingly for the detection outlier purpose.

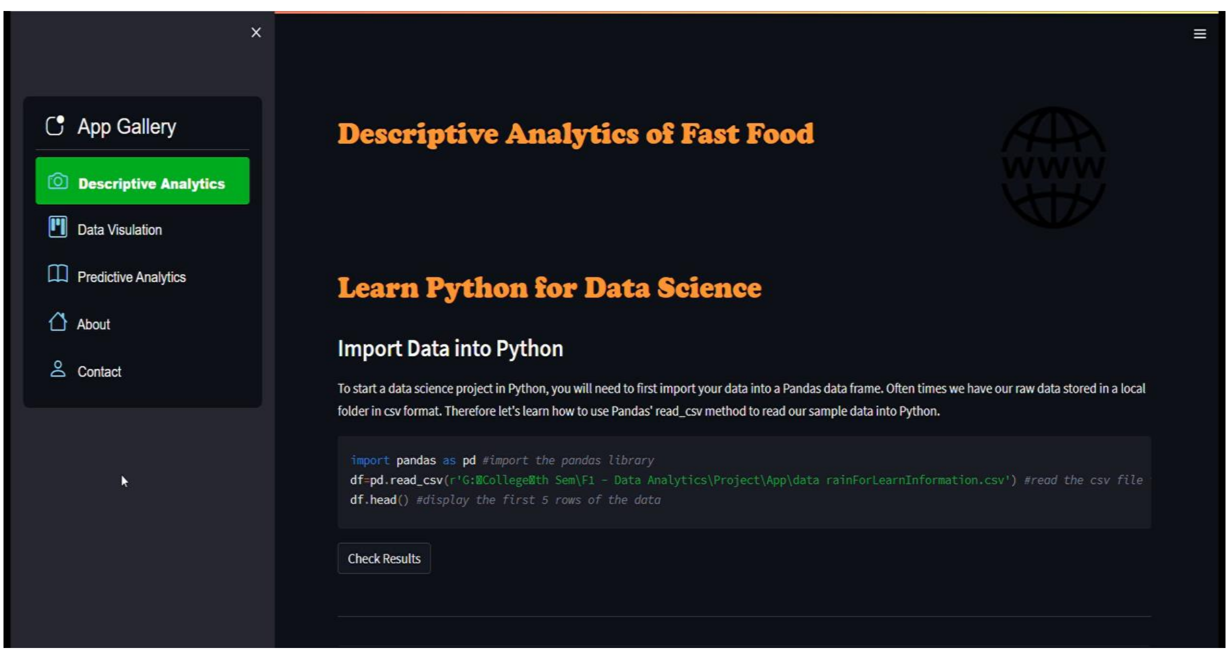
6) Dataset Splitting as Test set and Trains set

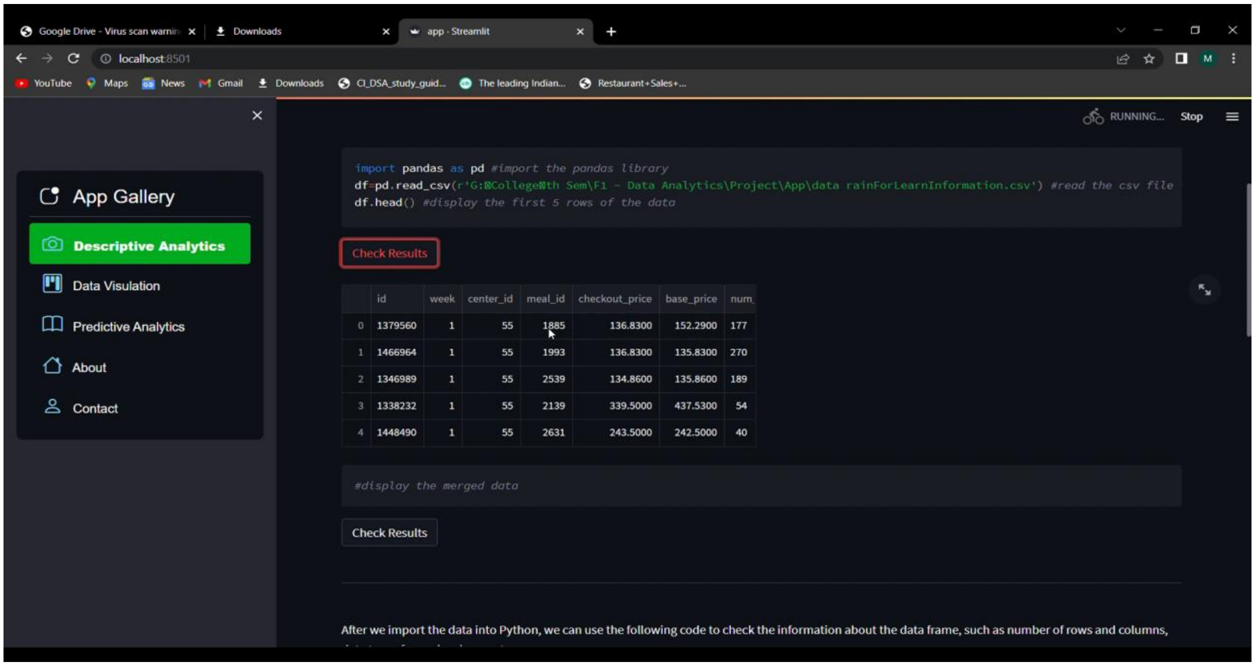
It was necessary to drop some variables that are not affecting the prediction to improve the prediction result. Such as variables "id" and “city_code” are identified as irrelevant variables for the train, “num_orders” is a target variable for prediction, “special price” variable calculation of base price and checkout price. But identified there is lack of correlation with the target variable, “week” variable categorized with quarter/year wise and “special price percent” also removed. After removing irrelevant variables, it was decided to fit catboostRegressor model to the training data using the fit method. Therefore, it was able to predict result based on this data using predict method .

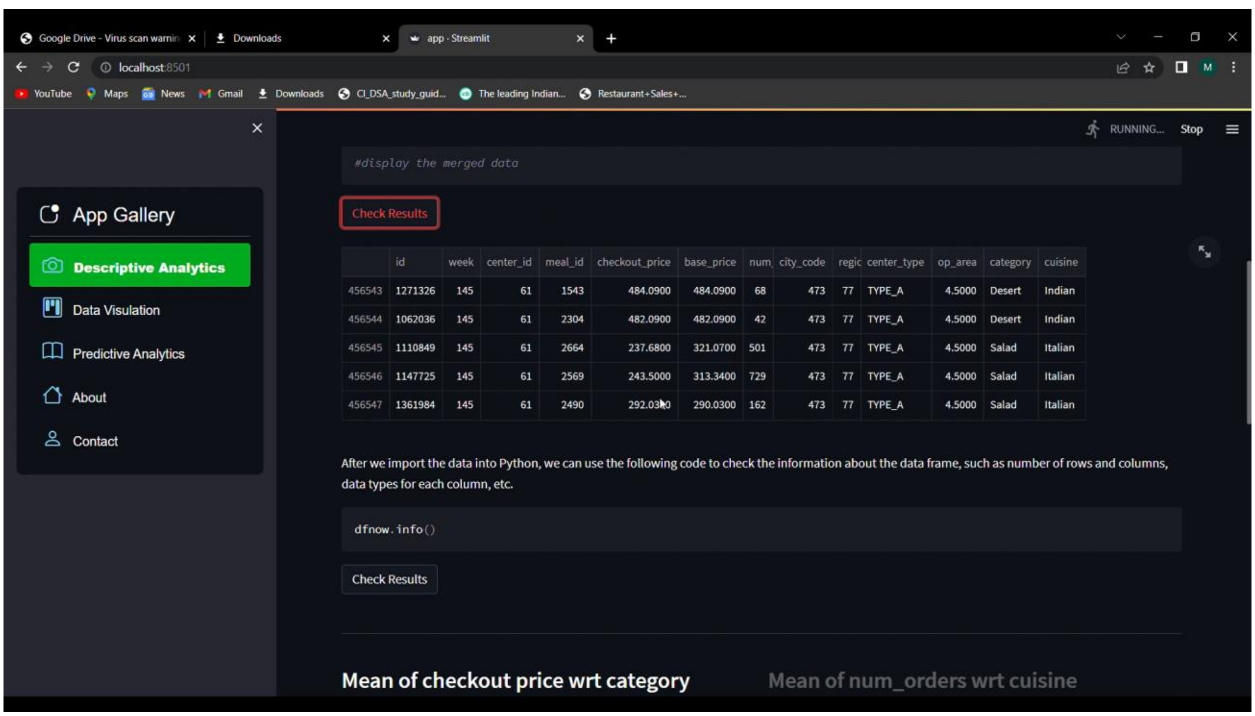
7) Model Training and Data Prediction

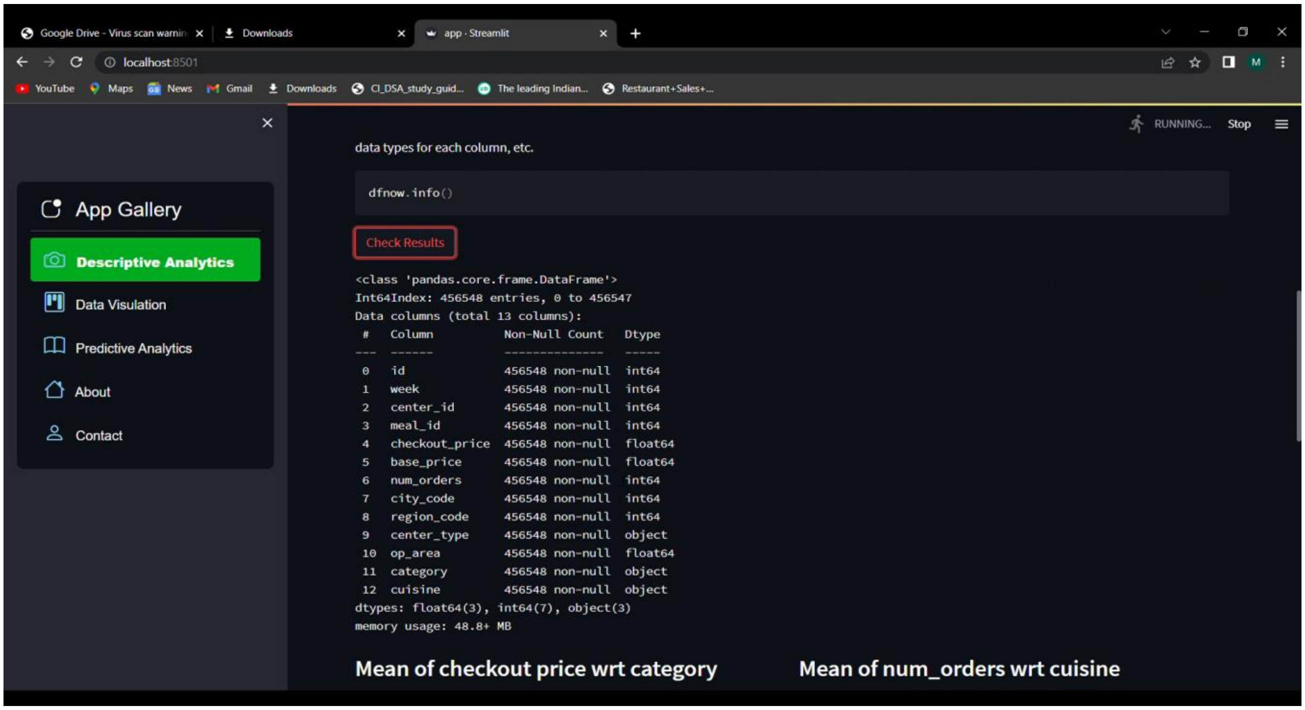
The predicted result was evaluated according to the implemented standard evaluation metrics .

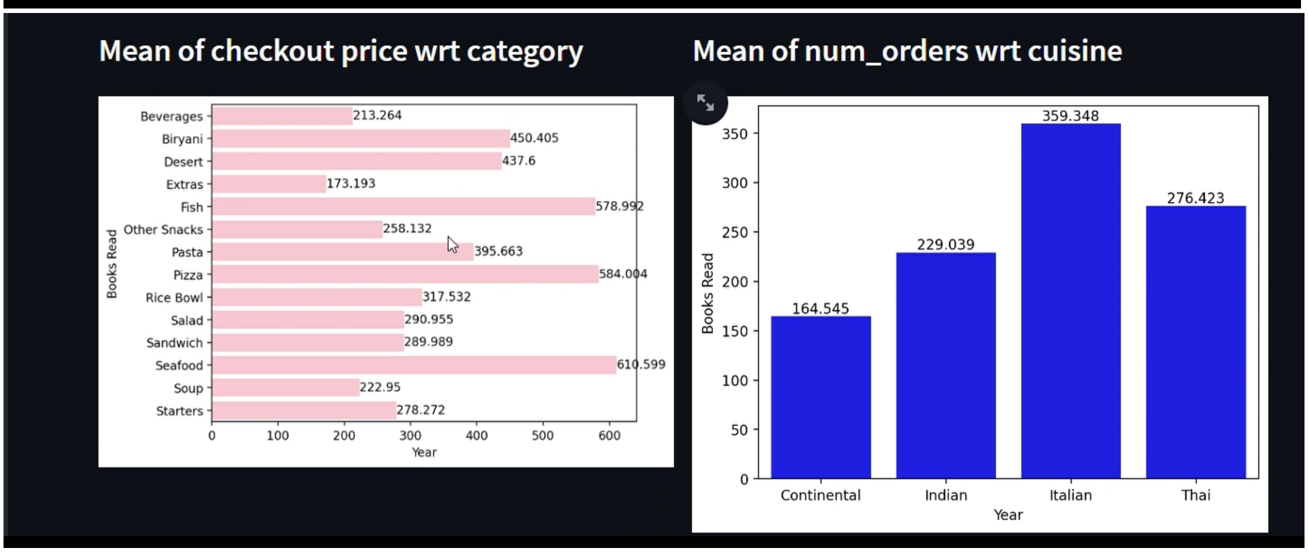
Implementation

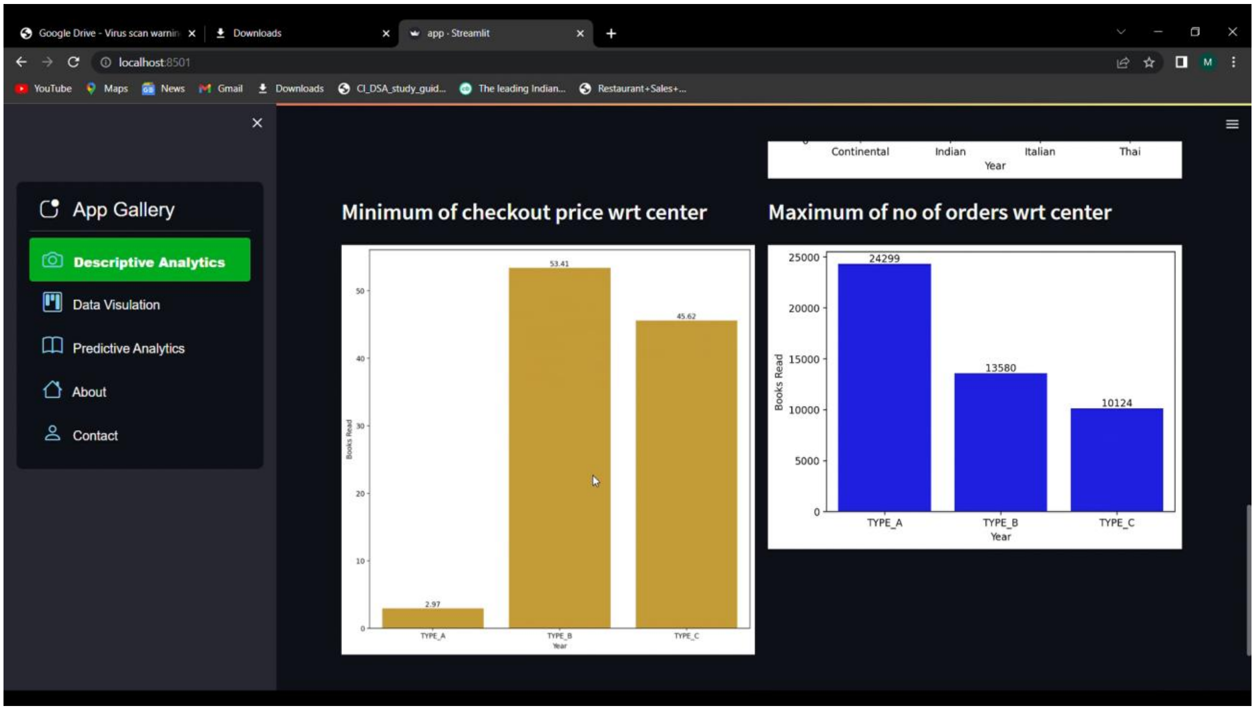



App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
RUNNING...
Stop
import pandas as pd #import the pandas library
df=pd.read_csv(r'G:\College\8th Sem\F1 - Data Analytics\Project\App\data rainForLearnInformation.csv') #read the csv file
df.head() #display the first 5 rows of the data
Check Results
	id	week	center_id	meal_id	checkout_price	base_price	num
0	1379560	1	55	1885	136.8300	152.2900	177
1	1466964	1	55	1993	136.8300	135.8300	270
2	1346989	1	55	2539	134.8600	135.8600	189
3	1338232	1	55	2139	339.5000	437.5300	54
4	1448490	1	55	2631	243.5000	242.5000	40
#display the merged data
Check Results
After we import the data into Python, we can use the following code to check the information about the data frame, such as number of rows and columns,

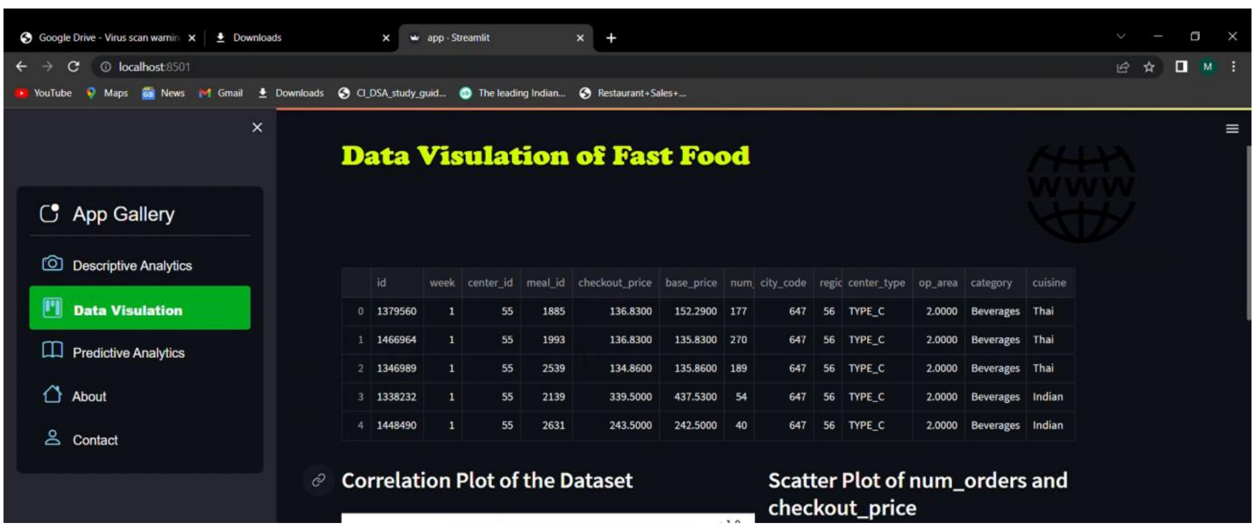

App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
RUNNING...
Stop
#display the merged data
Check Results
	id	week	center_id	meal_id	checkout_price	base_price	num	city_code	regic	center_type	op_area	category	cuisine
456543	1271326	145	61	1543	484.0900	484.0900	68	473	77	TYPE_A	4.5000	Desert	Indian
456544	1062036	145	61	2304	482.0900	482.0900	42	473	77	TYPE_A	4.5000	Desert	Indian
456545	1110849	145	61	2664	237.6800	321.0700	501	473	77	TYPE_A	4.5000	Salad	Italian
456546	1147725	145	61	2569	243.5000	313.3400	729	473	77	TYPE_A	4.5000	Salad	Italian
456547	1361984	145	61	2490	292.0300	290.0300	162	473	77	TYPE_A	4.5000	Salad	Italian
After we import the data into Python, we can use the following code to check the information about the data frame, such as number of rows and columns, data types for each column, etc.
dfnow.info()
Check Results
Mean of checkout price wrt category
Mean of num_orders wrt cuisine


Google Drive - Virus scan warnin
Downloads
app · Streamlit
localhost:8501
YouTube
Maps
News
Gmail
Downloads
CI_DSA_study_guid...
The leading Indian...
Restaurant+Sales+...
RUNNING...
Stop
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
data types for each column, etc.
dfnow.info()
Check Results
<class 'pandas.core.frame.DataFrame'>
Int64Index: 456548 entries, 0 to 456547
Data columns (total 13 columns):
Column Non-Null Count Dtype
0 id 456548 non-null int64
1 week 456548 non-null int64
2 center_id 456548 non-null int64
3 meal_id 456548 non-null int64
4 checkout_price 456548 non-null float64
5 base_price 456548 non-null float64
6 num_orders 456548 non-null int64
7 city_code 456548 non-null int64
8 region_code 456548 non-null int64
9 center_type 456548 non-null object
10 op_area 456548 non-null float64
11 category 456548 non-null object
12 cuisine 456548 non-null object
dtypes: float64(3), int64(7), object(3)
memory usage: 48.8+ MB
Mean of checkout price wrt category
Mean of num_orders wrt cuisine

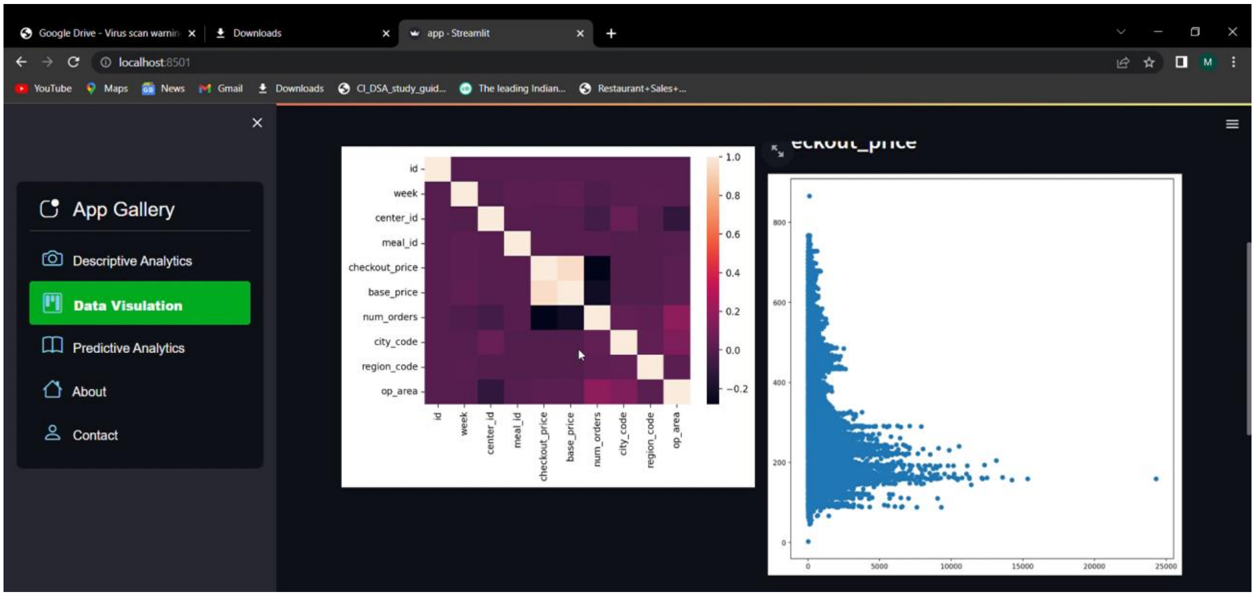

Mean of checkout price wrt category
Beverages 213.264
Biryani 450.405
Desert 437.6
Extras 173.193
Fish 578.992
Other Snacks 258.132
Pasta 395.663
Pizza 584.004
Rice Bowl 317.532
Salad 290.955
Sandwich 289.989
Seafood 610.599
Soup 222.95
Starters 278.272
Books Read
0 100 200 300 400 500 600
Year
Mean of num_orders wrt cuisine
Continental 164.545
Indian 229.039
Italian 359.348
Thai 276.423
Books Read
0 50 100 150 200 250 300 350
Year



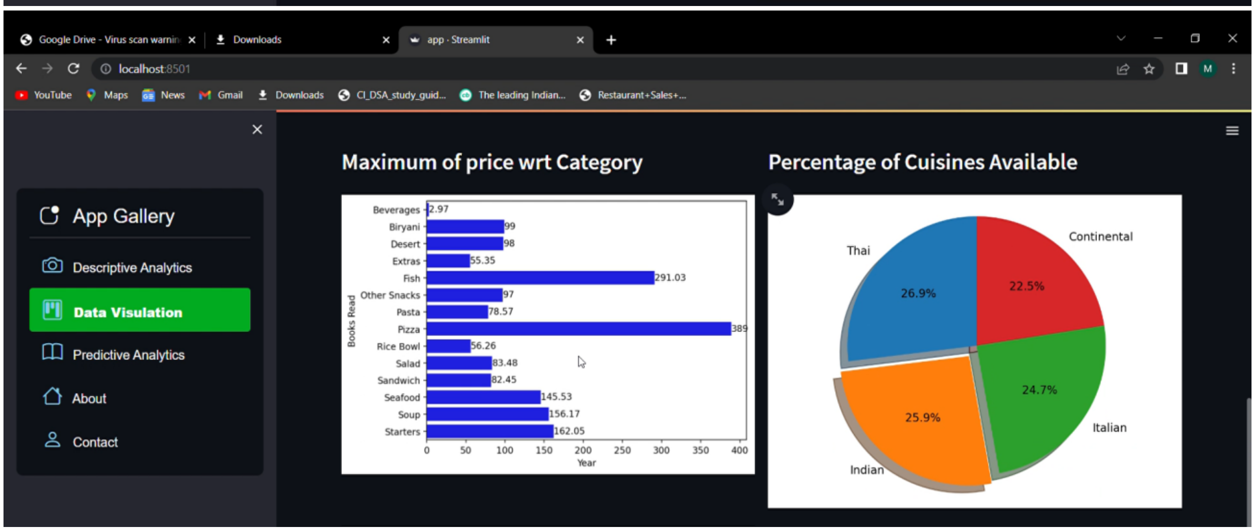
localhost:8501
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
Continental
Indian
Italian
Thai
Year
Minimum of checkout price wrt center
2.97
53.41
45.62
TYPE_A
TYPE_B
TYPE_C
Year
Maximum of no of orders wrt center
24299
13580
10124
Books Read
TYPE_A
TYPE_B
TYPE_C
Year



localhost:8501
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
Data Visulation of Fast Food
id week center_id meal_id checkout_price base_price num city_code regic center_type op_area category cuisine
0 1379560 1 55 1885 136.8300 152.2900 177 647 56 TYPE_C 2.0000 Beverages Thai
1 1466964 1 55 1993 136.8300 135.8300 270 647 56 TYPE_C 2.0000 Beverages Thai
2 1346989 1 55 2539 134.8600 135.8600 189 647 56 TYPE_C 2.0000 Beverages Thai
3 1338232 1 55 2139 339.5000 437.5300 54 647 56 TYPE_C 2.0000 Beverages Indian
4 1448490 1 55 2631 243.5000 242.5000 40 647 56 TYPE_C 2.0000 Beverages Indian
Correlation Plot of the Dataset
Scatter Plot of num_orders and checkout_price

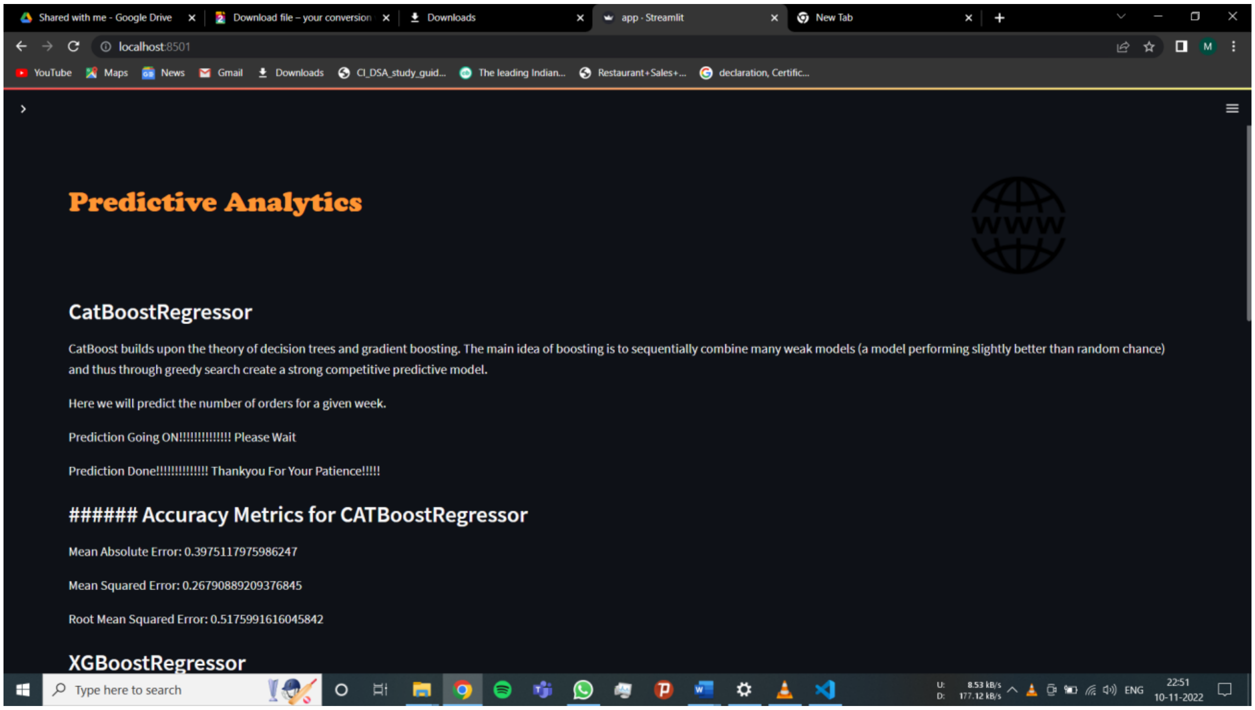


app · Streamlit
localhost:8501
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
eckout_price
id
week
center_id
meal_id
checkout_price
base_price
num_orders
city_code
region_code
op_area

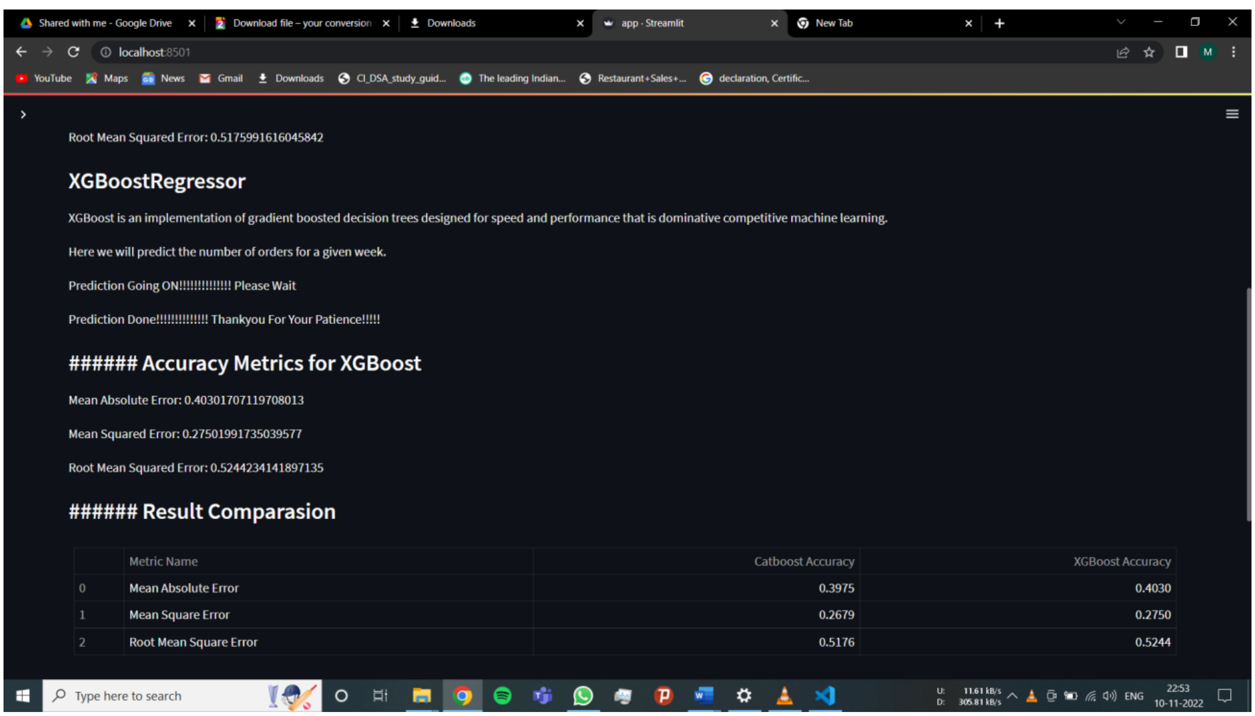


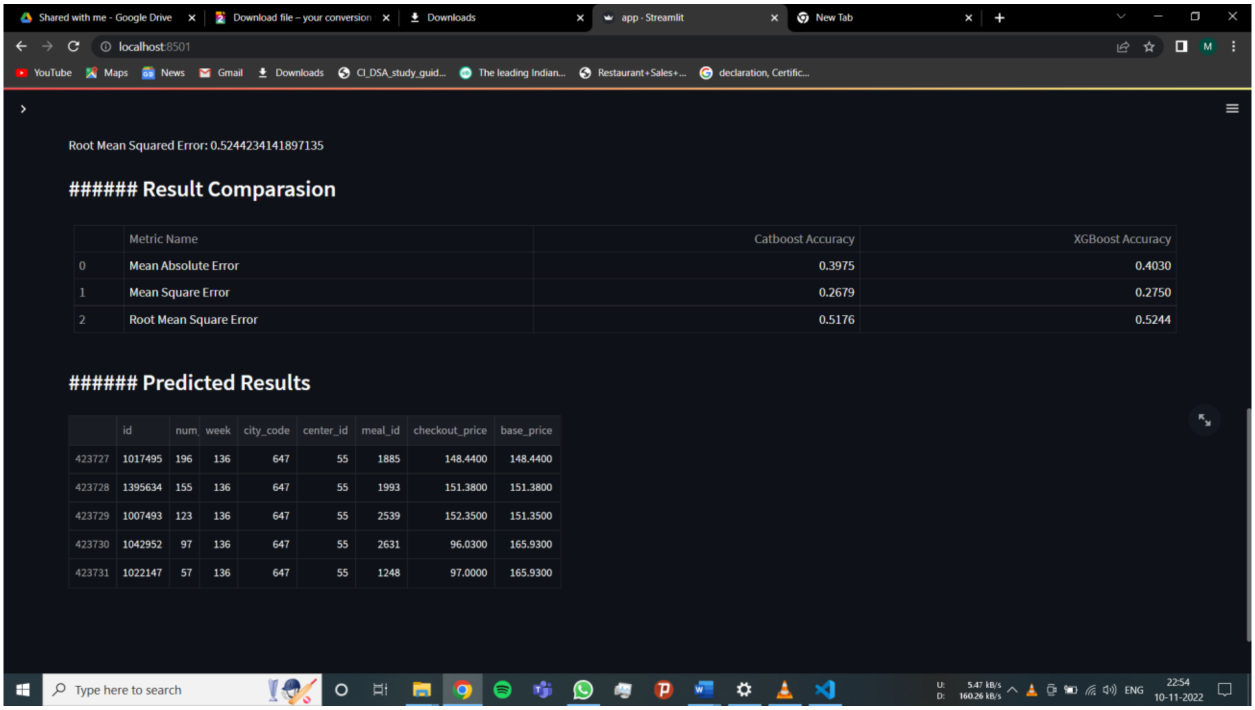
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
Maximum of price wrt Category
Percentage of Cuisines Available
Beverages
Biryani
Desert
Extras
Fish
Other Snacks
Pasta
Pizza
Rice Bowl
Salad
Sandwich
Seafood
Soup
Starters
Books Read
Year
Thai
Continental
Italian
Indian
26.9%
22.5%
24.7%
25.9%

Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
TYPE_B
TYPE_A
TYPE_C

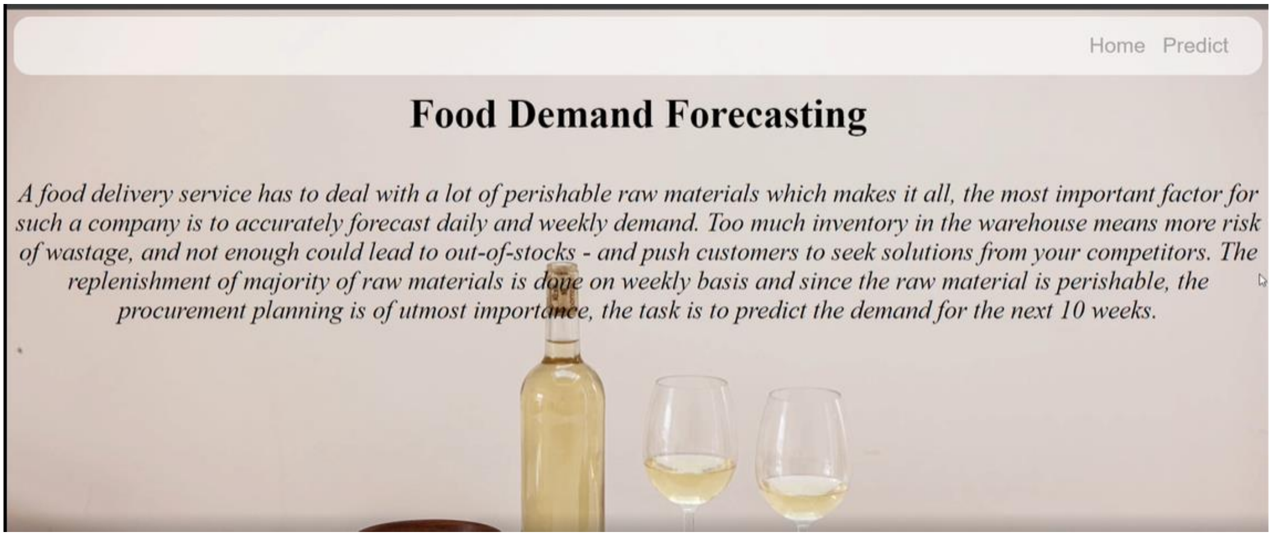


This is for cat boost regressor

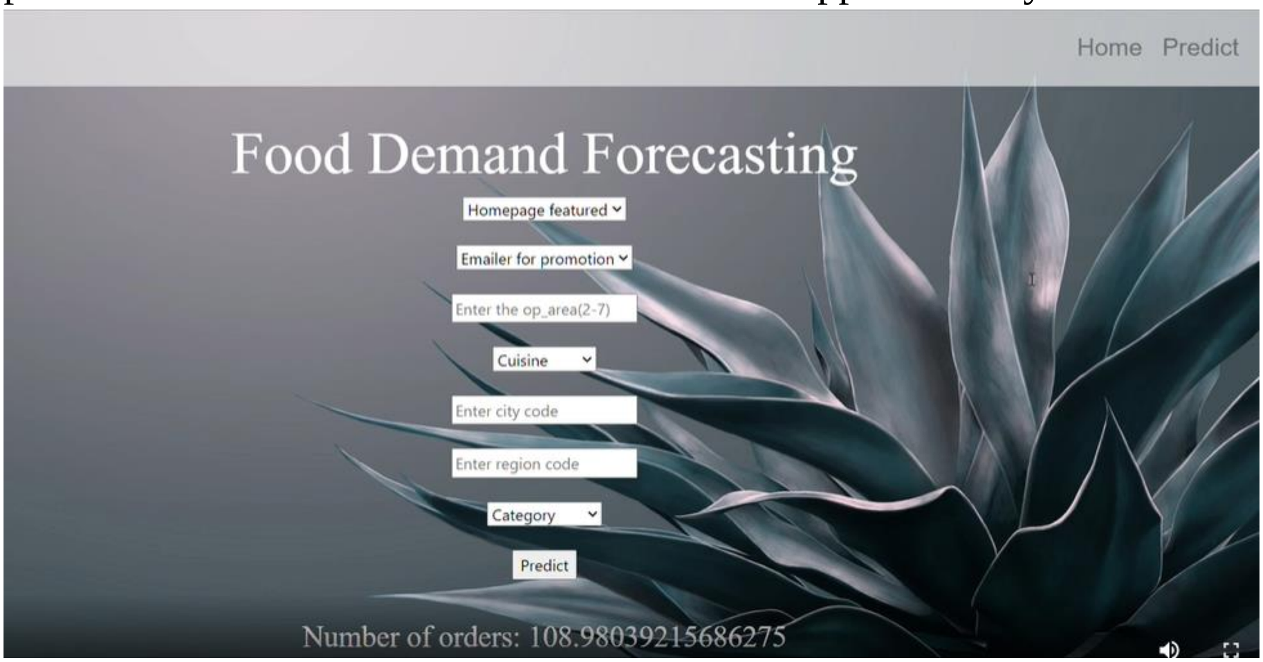


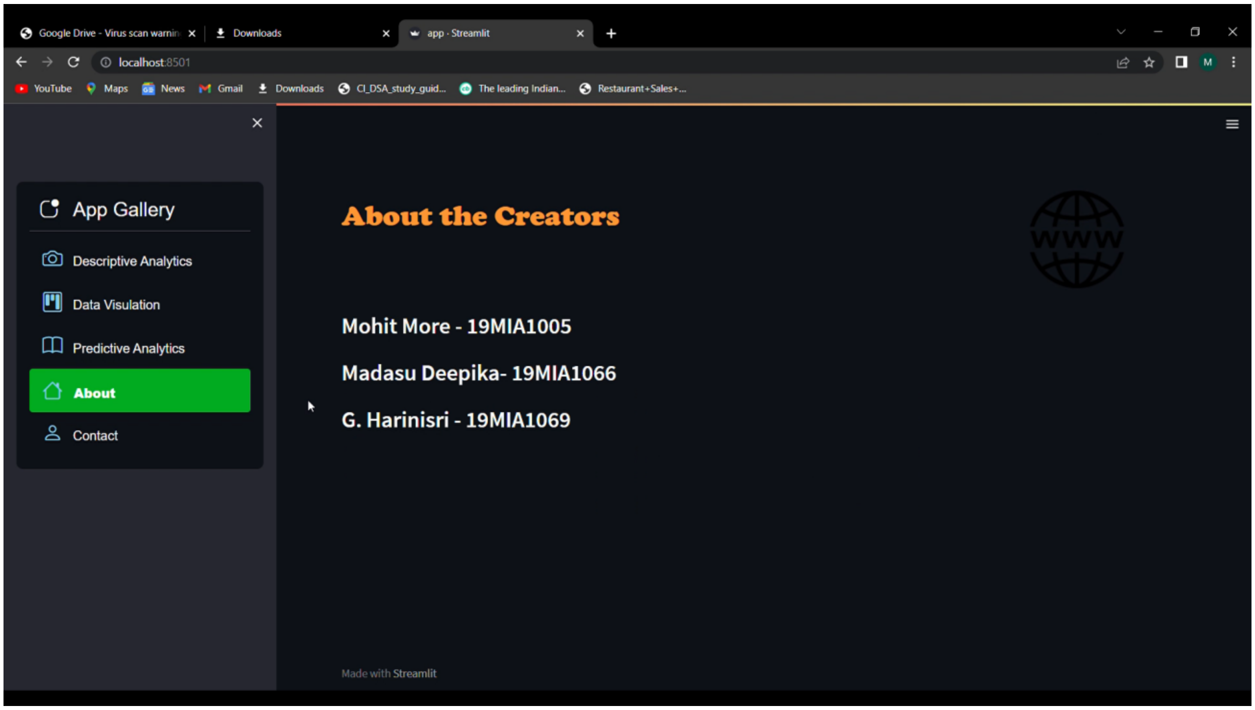


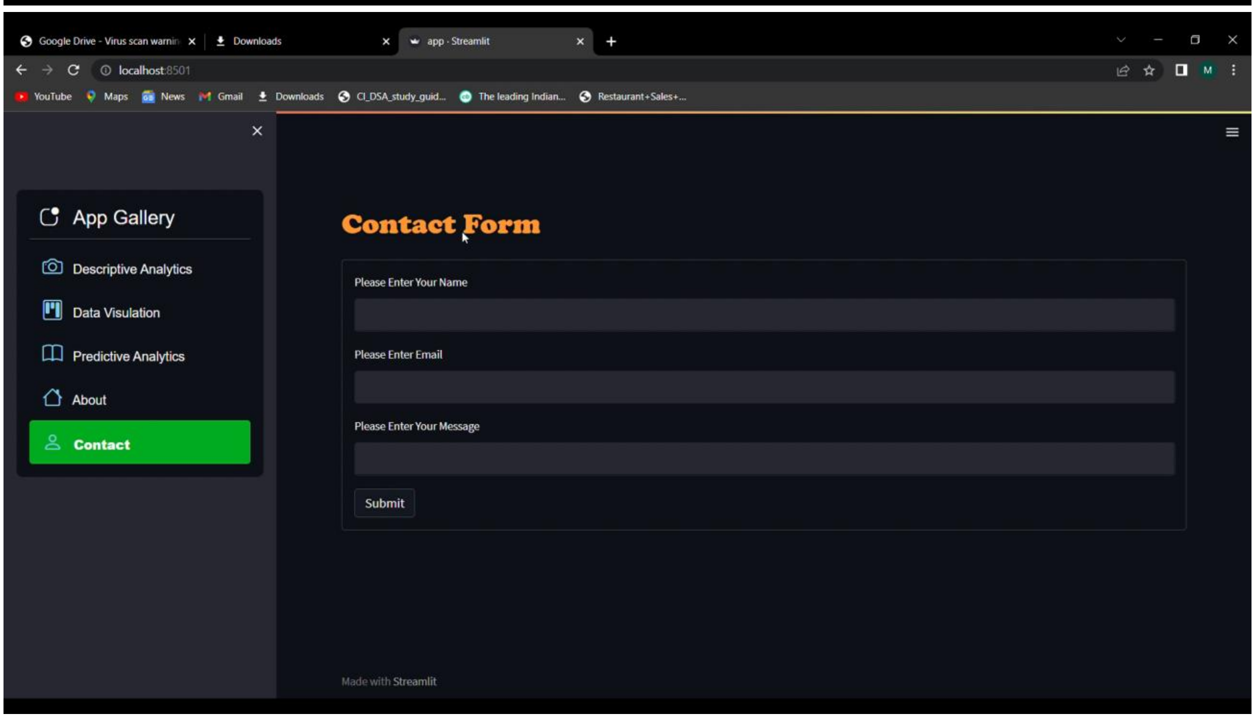
Also we have tried many machine learning algorithms like: LinearRegression, KNeighborsRegressor,Decision tree regressor , GradientBoostingRegressor. In which decision tree model has given the best accuracy .And we have prepared a prediction page which shows the number of orders according to the parameters given.



For the above given parameters like cuisine to be continental , op_area=2,city code=657,region code=56,category =biryani. The predicted number of orders is 108.98 which is approximately 109.




Google Drive - Virus scan warnin
Downloads
app · Streamlit
localhost:8501
YouTube
Maps
News
Gmail
Downloads
CI_DSA_study_guid...
The leading Indian...
Restaurant+Sales+...
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
About the Creators
WWW
Mohit More - 19MIA1005
Madasu Deepika- 19MIA1066
G. Harinisri - 19MIA1069
Made with Streamlit


Google Drive - Virus scan warnin
Downloads
app · Streamlit
localhost:8501
YouTube
Maps
News
Gmail
Downloads
CI_DSA_study_guid...
The leading Indian...
Restaurant+Sales+...
App Gallery
Descriptive Analytics
Data Visulation
Predictive Analytics
About
Contact
Contact Form
Please Enter Your Name
Please Enter Email
Please Enter Your Message
Submit
Made with Streamlit

CONTRIBUTION

Madasu Deepika – Front end and prediction

G. Harinisri – Descriptive Analytics and data visualization

Mohit More – Preprocessing , prediction algorithm, evaluation and comparison

Github and Website Link

https://github.com/mohitmore2001/Foundations-of-Data-Analytics-Project
https://mohitmore2001-foundations-of-data-analytics-project-app-ubvb4r.streamlit.app/

Conclusion

This research is conducted to evaluate the pattern of fast food and customer satisfaction encompassing various factors. The analysis of customer satisfaction encompasses independent variables which lead to consumer satisfaction and showcase the fast-food trend. Model training time and prediction time according to the following format, HH – represent Hour. MM – represent Minute. SS – represent Second. NS – represent Nano Second. It was certified that implementation of this model prediction accuracy very similar to the actual results.

The actual value tends to increase as the predicted values increases. Therefore, it is possible to say there is a linear positive correlation between those variables with a little number of outliers. At last, it was decided to use this model for the demand prediction process.

References

1. Tanizaki, Takashi & Hoshino, Tomohiro & Shimmura, Takeshi & Takenaka, Takeshi. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. Procedia CIRP. 79. 679-683. 10.1016/j.procir.2019.02.042.
2. Lutosławski, Krzysztof & Hernes, Marcin & Radomska, Joanna & Hajdas, Monika & Walaszczyk, Ewa & Kozina, Agata. (2021). Food Demand Prediction Using the Nonlinear Autoregressive Exogenous Neural Network. IEEE Access. 10.1109/ACCESS.2021.3123255.
3. Bozkir, Ahmet & Sezer, Ebru. (2011). Predicting food demand in food courts by decision tree approaches. Procedia CS. 3. 759-763. 10.1016/j.procs.2010.12.125.
4. Liu, Lon-Mu & Bhattacharyya, Siddhartha & Sclove, Stanley & Chen, Rong & Lattyak, William. (2001). Data mining on time series: An illustration using

fast-food restaurant franchise data. Computational Statistics & Data Analysis. 37. 455-476. 10.1016/S0167-9473(01)00014-7.

5. Hayashi, Yoichi & Hsieh, M-H & Setiono, Rudy. (2009). Predicting consumer preference for fast-food franchises: A data mining approach. Journal of the Operational Research Society. 60. 1221-1229. 10.1057/palgrave.jors.2602646.
6. K.Aishwarya1, Rao, A. N., Kumari, N., Mishra, A., & R, R. M. (2020). *FOODDE 07*(06). https://doi.org/International Research Journal of Engineering and Technology (IRJET)