Design/Practical Experience
Department of Computer Science and Engineering
Final Report

Academic Year : 2021-22

Semester: II

Date of Submission of Report: 01/05/2022

Name of Students: Gaurav Kumar (B19CSE032) & Harsh Kumar (B19CSE036)

Title of the Project : Predicting prices of crops in different Indian cities

Project Category: Summer/Winter/Semester projects with Institute faculty within or outside the department

Name of Mentors: Dr. Dweepobotee Brahma & Dr. Angshuman Paul

Targeted Deliverables:

- Data extraction and analysis
- Building of time series model for price prediction

Theory:

If we wish to predict the trend in prices of fruits or vegetables, time is an important factor that must now be considered in our models. A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. We have used the Autoregressive (AR) model and the Autoregressive Integrated Moving Average (ARIMA) model in this project.

Work Done:

Our first requirement was to extract data from the website which has records price details for different cities and crops. For this we used python programming language, selenium tool for data extraction and python libraries like pandas and numpy.We have downloaded chromedriver which is used by the selenium to control chrome and it helps to navigate between different web pages for chromium based web browsers.There are various sections such as Category Name and Crop Name which are to be filled which is done by the help of chromedriver.We used search by class, name and id to locate the required elements correctly on the website. After locating the required

data correctly we put them in a list. The dates for which there were no data available, we marked them as NA while putting them in the list. We ran a loop to extract the data for 6 months of the year 2021 and then added an extra column for the day. After putting all data in the list, we converted it into a pandas dataframe. After converting the data into dataframe, we convert it into a comma separated (csv) file so that it can be used in the final model.So, the code for the data extraction is attached below. After data extraction, we moved onto the data analysis part.

For manual analysis of data, we looked at all the data points manually and removed the outliers that is, those particular point values were much higher in comparison to others and the reason behind it was mistake in entry of data.Also, we removed all the points having NA values and then, we looked for variation in data to study the cost of crops per day. If there is no variation, then such data is of no use to us. For final model implementation, we have removed features like minimum and maximum prices as our main goal is to predict final retail price and retail price is inclusive of these features.

Finally, we applied the simple AR model on the dataset by taking first five months data for training and last month for testing our model.We also plotted the graph of actual and the predicted and plotted them together in a graph.Then, we extended the same concept to and extensive ARIMA model which takes into consideration both AR component and MA component.To find out the parameters, we plotted the acf graph and other graphs to calculate the value corresponding to them.Also, we plotted the final actual and predicted graph and study the results related to it.

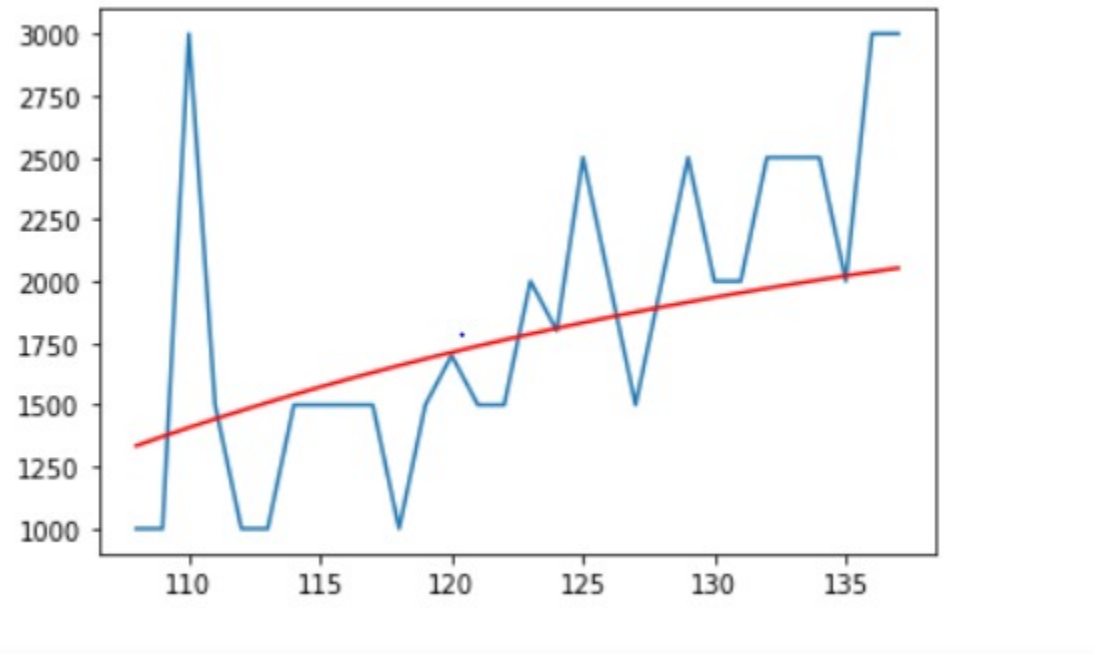
Future Work: We are looking to train these datasets on the more advanced models including both the AR and MA component known as ARIMA models which contain better ways to adapt to new datasets which have different crops in different cities.Our future includes running the advanced model such as ARIMA on these datasets extracted from the websites.

Sample Result-

Finally after running the data extraction part, we get the output in following format:

	Date	Min Price\n(RS/Qtl)	Max Price\n(RS/Qtl)	Model Price\n(RS/Qtl)	Arrival\n(MT.)	Retail Price\n(RS/Qtl)	City	Day
0	02/01/2021	2250	2600	2350	45	4000	CHANDIGARH	Friday
1	03/01/2021	NA	NA	NA	NA	NA	CHANDIGARH	Saturday
2	04/01/2021	3200	3600	3450	45	5000	CHANDIGARH	Monday
3	05/01/2021	2400	2800	2550	50	4000	CHANDIGARH	Thursday
4	06/01/2021	2000	2500	2350	40	4000	CHANDIGARH	Monday
...	...	...	...	...	...	...	...	...
170	26/06/2021	1400	1800	1560	70	3000	CHANDIGARH	Monday
171	27/06/2021	NA	NA	NA	NA	NA	CHANDIGARH	Thursday
172	28/06/2021	NA	NA	NA	NA	NA	CHANDIGARH	Monday
173	29/06/2021	1600	2000	1750	65	3000	CHANDIGARH	Saturday
174	30/06/2021	NA	NA	NA	NA	NA	CHANDIGARH	Friday

The output after running the dataset on the basic AR model:



AR model on tomato prices for Chandigarh (red line : predicted price, blue line : actual price)

Links:

Colab Notebook link of model implementation :
https://colab.research.google.com/drive/1Loz66rOOa4g4BQ0QH8QjFvnSclkms8OQ?usp=sharing

Script for data extraction:
https://drive.google.com/file/d/1c7i4pIyMy-GmKXocDJ6-eLUhACcqcjXe/view?usp=sharing