**Gramoday Price Prediction Machine learning**

**Sanjeev Kumar Khurana**

Hello, this is solution of the assignment. Please read it carefully.

**(b)** After plotting the graphs of prices of the three major markets in Agra, we observe a reoccurring pattern. We see that the graphs look very similar, indicating that the three markets function in a somewhat similar manner. We see that the maximum of all the three markets occur at around 3000-4000. All graphs fall downward after this maximum. And after that, for some time, all graphs flatten before taking another downward slide.

All in all, we conclude that the markets are working in a way which is very similar. The maximum value is almost similar in all markets indicating there is some behind-the-scene connection between these markets.

**(c)** For price prediction, I will use linear regression algorithm.

**i.** I will perform following data pre-processing on this dataset to make it useable for modelling:

1. Drop Sl.no. column. It is of no use.
2. Encode Market names as 0,1,2,… because we don’t use texts in modelling. After encoding Market names as 0,1,2… we will be able to use this column as one of our features.
3. Since we already know the District name, we can safely remove this column from the dataset. All values are Agra, so this column is not helping us in any way.
4. Commodity column is of no use as all values are potato there. So, I will drop it.
5. Just like what we did to “District name” column, we will also encode the “Variety” column. Again, encoding it with 0,1,2… will make it another useful feature because quality/variety of a particular item always affects its price.
6. All values in “Grade” column are FAQ, so we can again safely ignore this column. Drop it.
7. The next two columns, Min price and Max price are also useful features. Although if one wants to calculate just the model price, one can drop these two columns and safely ignore them. However, I will use these two features in my modelling. If we are not sure whether to drop them or not, we can calculate correlation coefficient between the three columns, Min Price, Max Price and Model Price. If there is some sort of similarity among these three features, this might cause problems in our model. Then we might have to drop the Min and Max price columns.
8. Model Price is of course our main feature. In technical terms, it is our “y” variable in the modelling since it is something we want to predict.
9. Price dates is also a very important feature, because prices might vary throughout the year depending upon the season and climate. Potatoes are grown in particular period, so maybe only for a few months, sales go high and prices might fall in these months due to excess supply.

**ii.** I will use following features: Market name, Variety, Price Date, and Min/Max Price (depending on the correlation coefficient)

**iii.**  This can be converted into a machine learning problem very easily. We will use the 4 aforementioned columns as our main features. The target variable will the Model Price. In technical terms, the four columns will constitute for “X” and Model Price will be “Y”. The objective of the algorithm will be to predict prices of potatoes.

**iv.** I will use linear regression. SVMs can also be used here, but I will prefer linear regression.

**v.** I will use Mean Square Error (MSE) as my loss function. It is the most commonly used regression loss function. It is given by sum of squared distances between our target variable and predicted values.

Formula is following: (It was difficult to type this, so I pasted a photo)

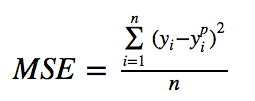


Image Source: Google

**Assumptions and Drawbacks**

Before signing off, I would like to point out that I have not paid much attention to the “Price Date” column. Although, in the python notebook, I have sorted the entire dataset by Price Date column, but after that, I have not taken that column into much consideration. However, since it is a very important column, during modelling, I will have to do some serious pre-processing with it. Also, during fetching the data process, I observed that due to some technical complications, it was difficult to get the whole data at once. So, I just fetched the data given on the first page, that is, I have not taken the data which was appearing by clicking at the “Next” button. However, this is just for showing you the way to get the data using web scraping tools like BeautifulSoup. If I had to do modelling, I will require the whole datasheet. This datasheet was available as an excel file on the website. So, for the sake demonstration of the solution, I downloaded it and loaded the data in the Jupyter notebook.

That concludes the assignment solution. I have attached the corresponding python script. Hope you find it. Hope you liked my solution. Waiting for positive response.

Thanks