

Name: Thomas Tran

Project Name: Store Sales - Time Series Forecasting

1. Introduction:

Kaggle hosts various challenges and competitions, the goal of this challenge is to use time-series forecasting to predict store sales of thousands of items sold at different Corporacion Favorita stores.

2. Dataset:

This dataset can be found on Kaggle; it includes a training set that contains location of store, type of product sold, total sales of product on a given date and total number of items being promoted on a given date. Also contains some other data including oil prices on a given day that may affect the economy of Ecuador. It also includes a test set to test the model on.

3. Data Preparation:

The first step in the data process is to clean and prepare the data. For this dataset, one thing that will need to be adjusted is changing the categorical variables by using dummy variables. Will also may need to adjust the dates to change into a continuous variable. Finally may need to remove some noisy features from both the training and test dataset. May also need to

4. Exploratory Data Analysis:

After the data has been prepared; can take a look at the data by viewing how the different variables are distributed and possibly if there is a correlation between any of the variables and sales on a given day. This may help determine a pattern or relationship between the explanatory variables and sales of a given item.

5. Modeling:

Will have to take a look at the data before making a concrete decision on the specific model that would be appropriate. However, I may use linear regression for time series forecasting. May also use quadratic regression or K-nearest neighbor to compare the accuracy of the different models.

6. Evaluation:

To test the accuracy of the models, may have to split the training set into 80% and 20% to test the accuracy of the models using F1 score or ROC curve. Alternatively, since it is a Kaggle challenge, it is possible to submit the submission to Kaggle and receive a score back.