R&A | DS Task | APMC/Mandi Machine Learning SocialCops Task

Ritwik M G

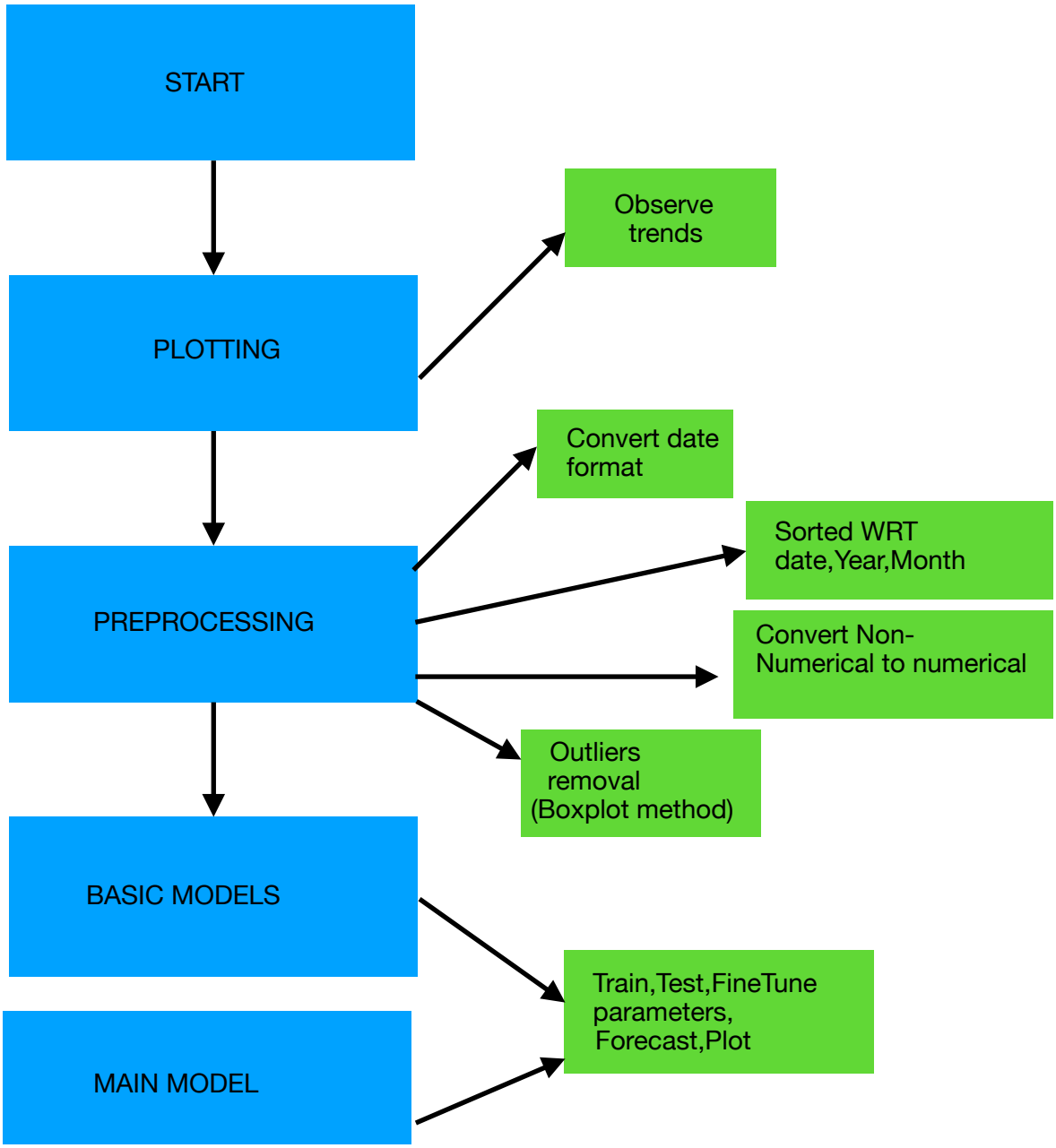
INDEX	
SR. NO	**CONTENTS**
1	File contents
2	Methodolgy
3	Assumptions

1. FILE CONTENTS

1. **predictXgboost.py**: First code attempt using the scikit learn library to better understanding the data and see its response against XGboost, RandomForestRegression.
2. **arima.py**: Full problem statement better handled by ARIMA model considering the seasonal and time series patterns.
3. **OutputPlotsPDF**: Consolidated PDF file having the main Output plot images for easy reference.
4. **OutputFilesPDF**: Consolidated PDF file having the screenshots of the output of predictions and accuracy obtained using various algorithms , with and without outliers removal.

5. ***OutputScreenshots***: Folder containing the output screenshots.
6. **Plots**: Folder contains the screenshots of the plots.

2. METHODOLOGY



Description:

1.**Plotting**: Data was uploaded to Meatabase(Open source data analysis platform) via syncing the data to PostGreSQL admin4 to get the basic plots and analyse the trends such as Highest,Avergae, Growth Decline etc among the data.

2.**Preprocessing**: Follwing steps were done to data before proceeding with the actual model training

- Append 'DD' to 'YYYY-MM' format of the given date so that plotting and feeding to Metabase is according to uniform standard format.

- Convert Columns like ‘Commodity’ to lowerCase so that there is no mismatch in the case while considering distinct types or any inconsistency for that matter.

- Convert Non-Numerical values containing columns to Numerical values so that it could be fed to the model. One hot representation was avoided to prevent Over-Fitting because Tree based methods were used.
- Outliers were removed using Box plot method by removing all the values beyond the Inter-Quantile range.
- Data was sorted according to Date,Year and then Month in sorted order so that it is fit to be as Index for Time series analysis.

3.**Basic models:** After the preprocessing plots were again analysed to see a more uniform data. State of the art Regression techniques like XGBoost and RandomForestRegression were used to fit on target features and near accuracy of 90% was achieved with both models and Crop prices of next three months were forecasted in accordance to this model.

4.**Main model:** ARIMA was chosen to be the primary model for solving such a problem because of the trends observed in the plots and such a type of the problem which includes seasonal trends and periodicity is best dealt with Time series analytical capabilities model like ARIMA.

Post the pre-processing step, date indexed based data was checked and modelled with the following :

- Check for stationarity using Rolling mean and Dickey-fuller test.
- Remove Trends from data using Exponentially weighted moving average method.
- Remove seasonality from the data by finding the type of seasonality either Additive/Multiplicative by evaluating the dependence between trend and Seasonal, residual components.
- Plotting ACF and PACF graphs to find the best fit Order for the ARIMA model ignorer to minimise on RMSE.
- Using Itertools and minimising AIC to see verify the best order for SARIMA.
- Fit ARIMA using the above order.
- Forecast the crop values for next three months.
- Plot the required graphs meanwhile.

3. ASSUMPTIONS

- Day of ’01’ was prefixed to the ‘YYYY-MM’ date for uniformity in data for visualisation purpose assuming the every months day was of 1st of the respective month.
- Each combination of APMC and Commodity is unique and hence trends of one combination cannot be assumed to prevail in another subset of combinations. Ex: Commodity Bajri would have different trends and significance in two different APMCs hence once’s data is totally useless for prediction of other combination.
- Due to limited computation power, the program is designed to compute and predict values of any one particular User required Commodity and APMC instead of calculating for each of such many combinations(349*352). Assumed due to the fact that the same computation logic can be iterated through each combination to get the result.
- Assumption in output: For the ARIMA model the scaled down version’s predicted values are unequal but very close, this is due to the fact that each of the above mentioned combination yield very less data individually(Approx. 25-30 records). hence when these close values are being scaled up the final prediction values are seeming to be equal, while actually they might not be.
- For predicting the Trends such as Best//worst profitable crop,Best/worst ROI(Return on investment) , which area/district yield most profit for a particular Commodity or in a season and

many other such trends can be predicted with the help of MSP(Minimum support price) but in this particular problem statement where just the future prices have to be forecasted the “CMO_MSP_MANDI.CSV” file was left unused.