Store Sales - Time Series Forecasting

I) Goal of the Project:

It uses time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer to build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores.

II) Datasets Description:

In the Dataset, we have data subsets as train, test, holidays_events, oil, stores, & transactions respectively.

- The training data comprises a time series of features **date**, **store_nbr**, **family**, **onpromotion**, and the target **sales**.
- The test data has the same features as the training data, except the target column i.e. sales.
- Holidays_events data consists of **date**, **type**, **locale**, **locale_name**, **description**, **transferred**.
- Store data includes **store_nbr**, **city**, **state**, **type**, and **cluster**.
- Oil data contains column **dcoilwtico**, which is the daily oil price.
- Transactions data contains transaction details like **date**, **store_nbr**, and **transactions**.

III) Data Field Information & additional notes.

- **store_nbr** identifies the store at which the products are sold.
- **family** identifies the type of product sold.
- **sales** gives the total sales for a product family at a particular store at a given date. Fractional values are possible.
- **on-promotion** gives the total number of items in a product family that were being promoted at a store on a given date.
- **cluster** is a grouping of similar stores.
- In holidays_event the transferred column, A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday.
- Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend).
- Additional holidays are days added to a regular calendar holiday.
- Wages in the public sector are paid every two weeks on the 15th and the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first-need products which greatly affected supermarket sales for several weeks after the earthquake.

IV) Data Preprocessing:

Understanding each column of all the data subsets, & relevant to the concern problem.
We have checked there are no missing/nan values in the entire dataset, but the oil_price i.e. **dcoilwtico** column should be a continuous price, but it wasn't, so we are filling the missing values through linear interpolation.

Now we are processing the holidays_event dataset, transformation for the Holiday that is transferred and should be updated to a normal day, Days that have type 'Transfer' will be updated as regular holidays. Now transformation has been done, we can drop the column "transferred", alongside the if the type of the holidays_event dataset has 'Bridge', we are updating to Holiday.

We are combining oil, holiday, transactions, and store datasets to the training dataset to have an entire training dataset while merging and filling the missing values as 'None'.

Turning the date column into a datetime object, so that we can do transformations more easily for the entire merged_train datasets, adding a "**payday**" column to the merged_train dataset, which reflects the binary value 1, if the date on which salary is paid, i.e on 15th & the last day of the month, otherwise 0. Drop the **last_day_of_month** column after transformation.

Due to the 7.8 earthquake that struck Ecuador on April 16, 2016, we have to drop the disaster date from the training data, so the ambiguous values will not affect the model training.

Combining oil, holiday, transactions, and store datasets to the test dataset to have an entire merged_test dataset, similar to the method used in obtaining the merged_train dataset. Similarly accounts for paybacks & disasters.

V) **Feature Selection:**

Identify the features that are most relevant for predicting sales. As we have replaced the missing values as 'None', while merging datasets, now for the object Dtype columns the values behave as an 'str', while for the nonobject columns like integer or float DType columns the values also behave as an 'str', so for the later one we have to replace it with the median of the corresponding columns, & for the object Dtype columns we are replacing with 'np. nan':

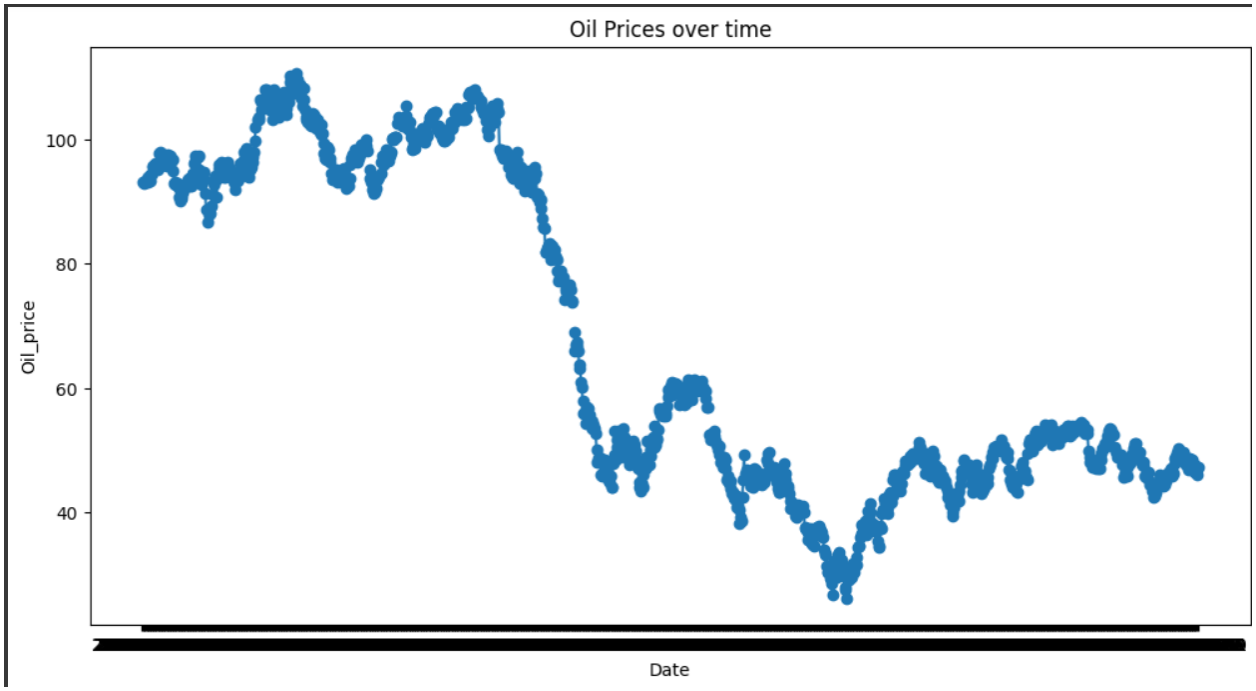
Applying label_encoder to object Dtype columns, to transform the categorical values to numerical values.

Computing correlations between the columns & the target column i.e. sales, to reduce the number of less relevant columns. [For merged_train dataset].

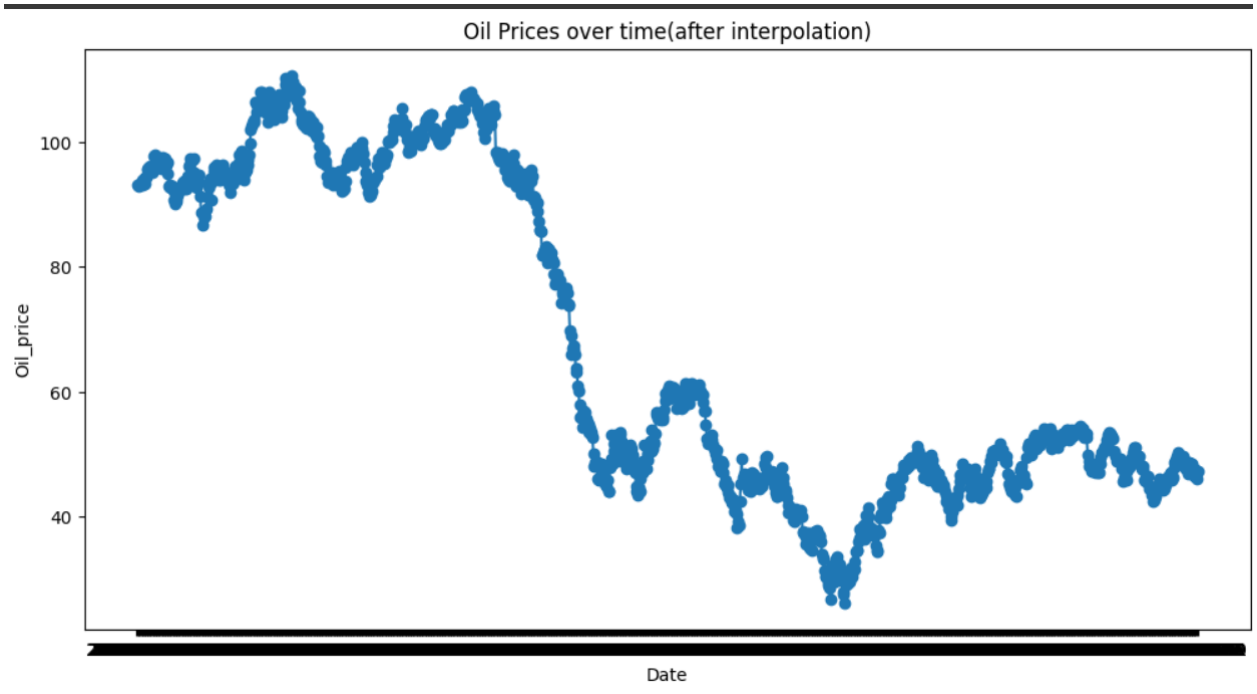
For the Final merged_train Dataset, extracting features based on the correlation values.

Repeating the same transformation to obtain relevant features for the merged_test dataset.

VI) Visualization/ Plots:



After Linear interpolation.



VII) Model Selection: Considering the entire merged_train dataset with the selected features and splitting the dataset into training and testing, further splitting the dataset training, validation & testing datasets, We opt for the Linear, Lasso, Ridge Regression, and then Random forest regressor.

VIII) Model Training: Train the selected models on the training data and fine-tune their hyperparameters for optimal performance. Monitor for overfitting and use techniques like cross-validation to assess model generalization.

IX) Model Evaluation: Evaluate model performance using appropriate metrics. Metrics scores like mean squared error (MSE) mean absolute error (MAE) or F1 score. So the model with a good metrics score can fit the dataset to predict the next sale values, we are imputing various hyperparameters of random forest regressor, to have a better result for selected sets of hyperparameters as `n_estimators`, `max_depth`, & `min_samples_split`.

Similarly for other models, like for Lasso, & Ridge regression, manually changing the alpha value, for the Support Vector Regressor adjusts the kernel and C parameter.