Sales Forecast – Data Wrangling & Visualization:

Sunday, October 22, 2017

1. Import packages for data analysis/wrangling
2. Import packages for data visualization
3. Read TRAIN and TEST csv files as dataframes
4. Date column in the TRAIN and TEST dataframes are object datatypes. Change this to datetime.
5. Identify the target variable from the business problem. For this problem, Weekly\_Sales is the target variable.
6. Read the Features and Stores csv files as dataframes.
7. Merge the Features and Stores dataframe into the train dataframe.
   1. Check which columns are available in both dataframes (train- features and train-stores) and use LEFT join to retain all records in train dataset.
8. TRAIN\_DF
   1. Inspect the head and tail of train\_df
      1. This should now include the fields from features and stores csv files.
   2. Check info() for train\_df
      1. Check if any of the columns in train\_df contain BLANK, NULL or EMPTY values, by looking at the info() count values.
   3. Check describe() for train\_df
   4. Create Year and Month columns from the DATE column.
   5. Create a new column and store the log transform of Weekly\_Sales in it.
   6. Define a function that will calculate the temperature band and create a new dataframe column and apply the function to this new column.
   7. Define a function that will calculate the CPI range and create a new dataframe column and apply the function to this new column.
   8. There are so many NaNs found in Markdowns1 through Markdowns5. Need to know if there is any use in substituting these NaNs with values.
   9. Analyze the distribution of numerical values, using value\_counts() on the numerical features.
      1. Number of stores
      2. Number of departments
      3. Number of records in each year
      4. Record count by store
      5. Record count by dept.
      6. KDE Plot of target variable – WeeklySales
   10. Analyze the distribution of categorical features
       1. Plot count of IsHoliday across years.
       2. Sales for all stores across all 3 years - monthwise, with isholiday flag.
       3. Use describe(include=[‘O’]) for categorical values.
       4. Use nunique, value\_counts() as needed.
   11. Visualize data
       1. Find top 5 selling stores
       2. Plot sales for all stores for 3 given years
       3. Plot sales for all stores for 3 given years, taking into account the ‘IsHoliday’ flag.
       4. Plot sales for store 1 and year 2010 (what does the x-axis represent here?)
       5. Plot sales for store 1, year 2011 – for 12 months – The graphs when plotted using hist, bar, line look very different; and what does the x-axis represent here? Can’t understand how data is plotted here.
       6. Sales for specific Store 20 for 3 years.
       7. Sales for specific 5 stores for 3 years.
       8. Sales of Store 1 over 3 years, by month
       9. Sales of all stores by year, month and holiday
       10. Sales of all stores by year, month and Type (of store)
       11. Sales of all stores by year, month and Temperature Band
       12. Sales of all stores by year, month and CPI range
       13. LM Plot (how to interpret this graph)