Summary of "Predicting Sales of every product and store on November 2015"

Kaggle Competition

Seungjun (Josh) Kim

https://www.kaggle.com/c/competitive-data-science-predict-future-sales

Given Files

1. sales_train.csv - the training set. Daily historical data from January 2013 to October 2015.
2. test.csv - the test set. You need to forecast the sales for these shops and products for November 2015.
3. sample_submission.csv - a sample submission file in the correct format.
4. items.csv - supplemental information about the items/products.
5. item_categories.csv - supplemental information about the items categories.
6. shops.csv - supplemental information about the shops.

Data Fields

7. ID - an Id that represents a (Shop, Item) tuple within the test set
8. shop_id - unique identifier of a shop
9. item_id - unique identifier of a product
10. item_category_id - unique identifier of item category
11. item_cnt_day - number of products sold. You are predicting a monthly amount of this measure
12. item_price - current price of an item
13. date - date in format dd/mm/yyyy
14. date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
15. item_name - name of item
16. shop_name - name of shop
17. item_category_name - name of item category

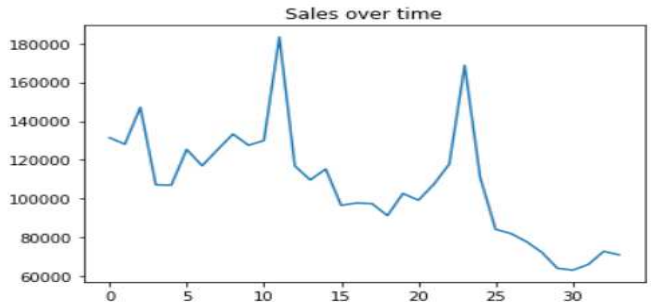
Methodology & Analysis

☞ Combined four datasets(sales_train, items, item_categories, shops) into one

☞ Dropped duplicates

☞ Checked for unreasonable values: The "item price" column had unreasonable values (excessively big or minus values). Dropped such values.

☞ Exploratory Data Analysis(EDA)



Mainly looked at the following three aspects for EDA:

- Monthly Sales over time

- Monthly Sales of different shops over time

- Monthly sales of different item categories over time

→ Sales peaked at the end(October-December) of each year probably due to end-of-year holiday season

☞ Feature Engineering

Created new features

Date	year	monthly sales	monthly sales mean	Item Description Length	Item Description Word Count
Length of Item Category Description	Item Category Description Word Count	Length of Shop Name	Shop Name Word Count		

☞ Various Machine Learning Models

```
col = [c for c in train.columns if c not in ['item_cnt_month','ID']]

#Validation
x1 = train[train['date_block_num']<33]
y1 = np.log1p(x1['item_cnt_month'].clip(0.,20.))
x1 = x1[col]

x2 = train[train['date_block_num']==33]
y2 = np.log1p(x2['item_cnt_month'].clip(0.,20.))
x2 = x2[col]
```

Model	**Validation RMSE**	**Actual RMSE at Kaggle Score Calculator**
Linear Regression	0.400294496205	2.52974
Passive Aggressive Regressor	0.469601300906	1.15786
Decision Tree Regressor(max_depth=3)	0.313631350404	1.58800
ExtraTreesRegressor(n_estimators=25, n_jobs=-1, max_depth=15)	0.298586931919	1.38177
Lasso LARS(alpha=0.01)	0.437092029485	2.03110
SGD Regressor	7.81039432528	6.49673
AdaBoostRegressor(tree.DecisionTreeRegressor(max_depth=3), n_estimators=100)	0.335047754266	1.71008

→ Passive Aggressive Regressor performed the best among the seven models tested

☞ Limitations

- More room for feature engineering

- Haven’t really used time-series modeling techniques including ARIMA