Applied Artificial Intelligence -Project5

Username-Toothless

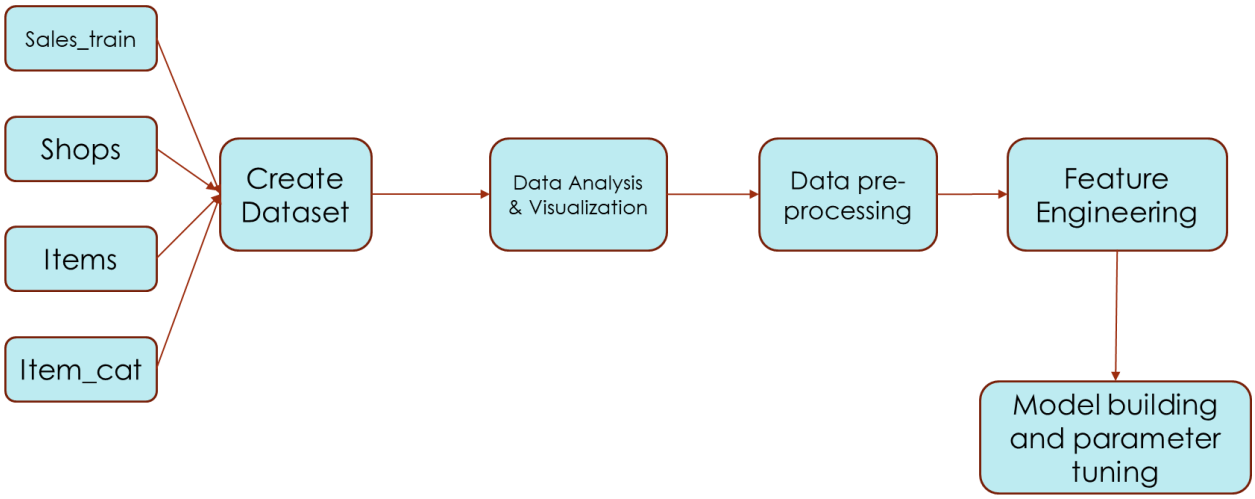
Domain:Kaggle -predict future sales

Abstract:

In this project we will work with a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.

We will predict total sales for every product and store in the next month.

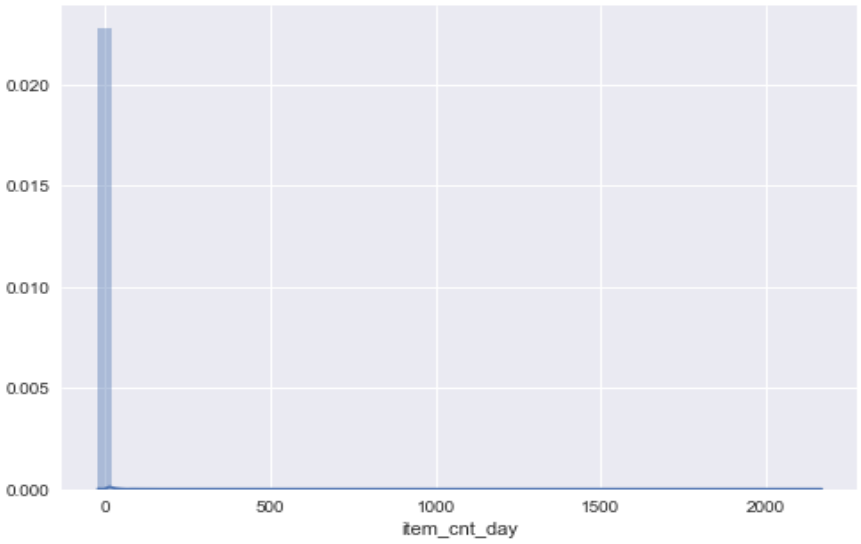
Flow diagram of modelling:



Explaining each block of flow diagram below:

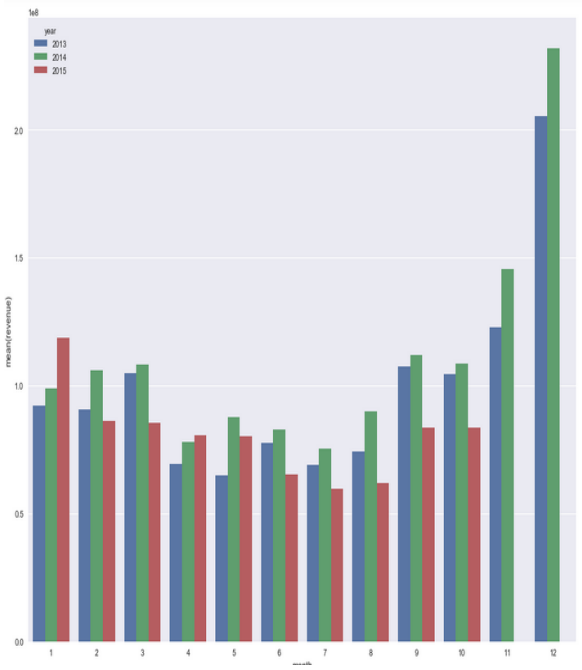
Dataset:

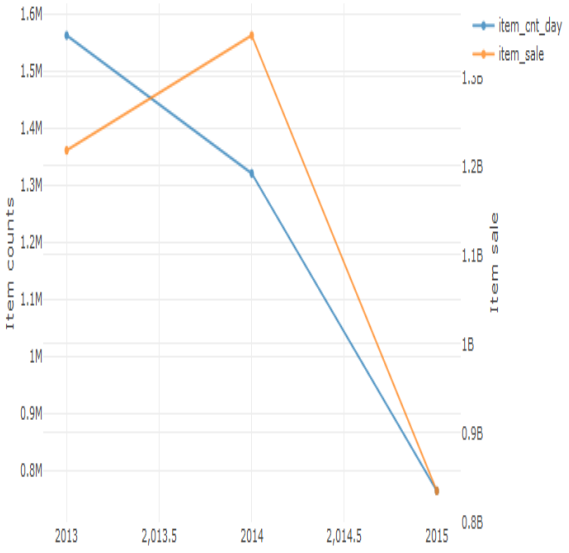
- We have **Time series** dataset: a dataset that has values over a period of time
- **Data fields**
- item_cnt_day - number of products sold. we are predicting a monthly amount of this measure
- date - date in format dd/mm/yyyy
- date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
- Target is heavily skewed, most of the values are zero. 0,20 range contain 99.9% of the data range.



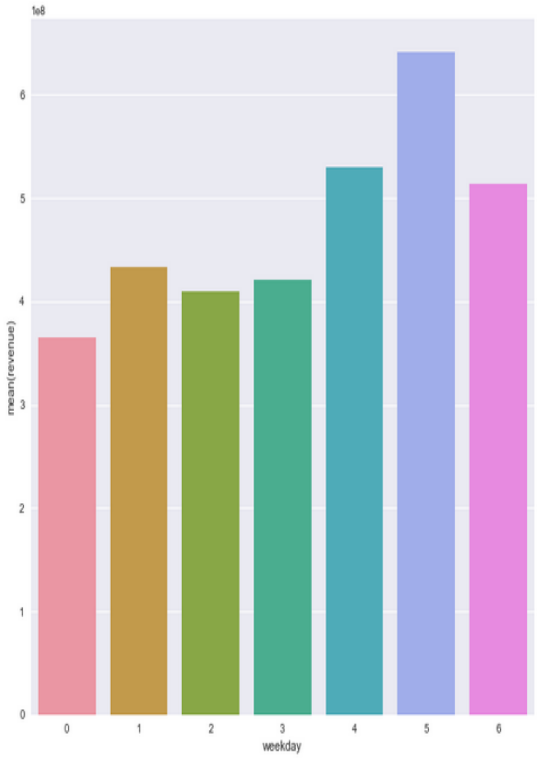
Data Analysis:

- Month & revenue plot





- Week and revenue plot:



There is a clear weekly cycle, with more sales on Fri and Sat

Decline in the revenue on Sunday can be accounted as most shops declare holiday.

Feature Engineering:

- Remove outlier and duplicates from sales_train data
- Calculate aggregation features for each month on shop_id and item_id, shop_id only ,item_id only, category_id only.
- Split the date column into month and year features
- Generate lag features for month 1or more
- For linear regression, only numerical features are fed into the model.
- Features encoded: item_id,shop_id,item_category_id,month,year
- Generated Revenue data columns by calculating product of item_cnt_day and item_price
- Validation-Train test split is time based used last two month as validation set
- Metrics optimization-Regressors minimize mean squared error. Validation metric used RMSE, same as the evaluation metric of the project.
- Hyperparameter tuning-used grid search to do parameter tuning for xgb and Lgbm

Model Comparison and Results:

- **Light Gbm:**

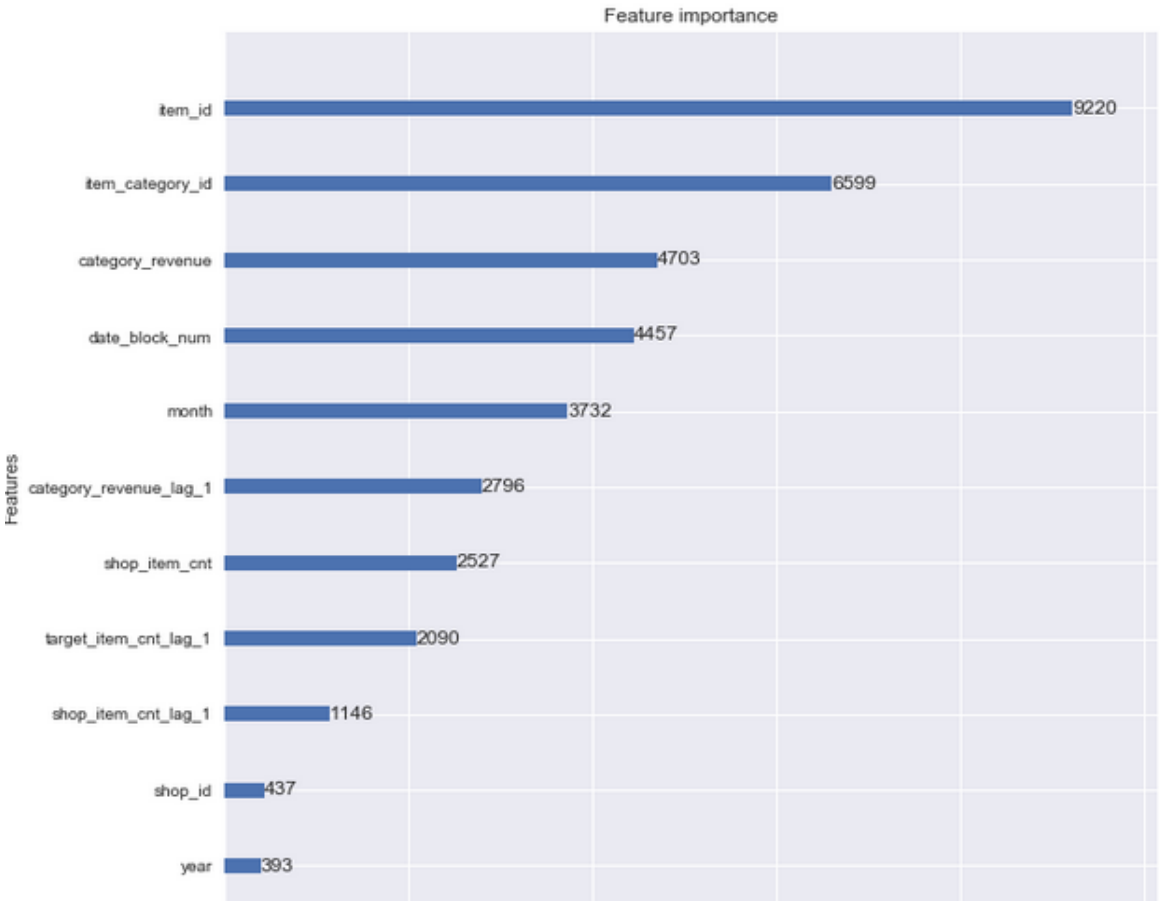
Results:

Train RMSE for lightgbm is 0. 822324

validation RMSE for lightgbm is 0. 929103

lightgbm runs for 144.32 seconds.

Feature importance for LGBM:



`result:

submission.csv

Future Scope:

- Generate more feature related to holiday, such as: differences between current month and holiday month.
- Convert text to data using :TFIDF vectorizer or word embedding

- Generate lag features with time steps:1,2,3,5 and 12 months.
- Use stacked architecture to improve score
- Use deep learning techniques-RNN(LSTM or GRU cells)

Tools used: