PREDICT FUTURE SALES

CMPE 255

Bijayani Sanghamitra Mishra - 014637753
Chaitrali Ajay Joshi - 014605149
Sreeja Madanambeti – 014619930

OBJECTIVE

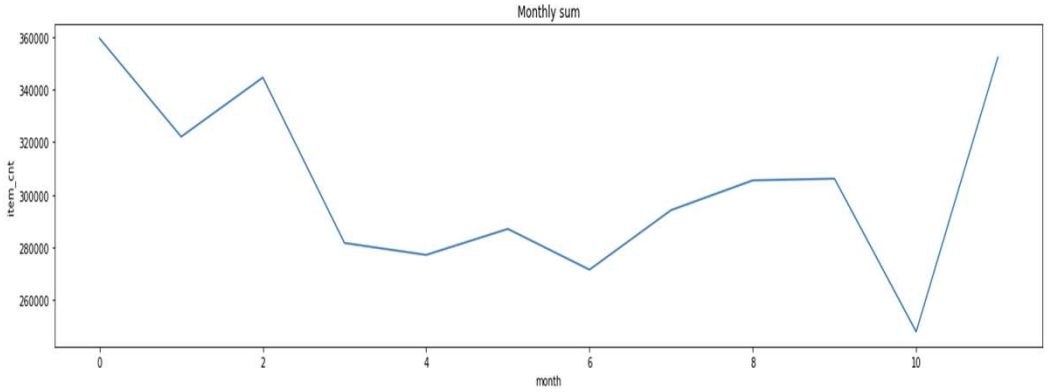
- For one of the largest Russian software firm "1C Company", we aim to forecast the total sales for every store and every product for one month.
- We have data available from January 2013 to October 2015. This project will focus on utilizing a variety of models to predict sales.

DATASET

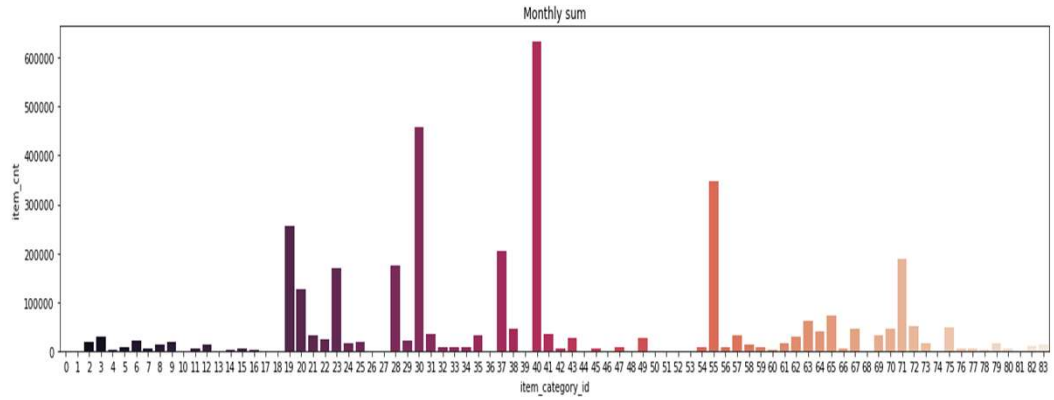
SOURCE	DESCRIPTION
sales_train.csv	Daily historical data from Jan 2013 to Oct 2015
items.csv	Items/products information
item_categories.csv	Items categories information
shops.csv	Shops information
test.csv	We Require to forecast the sales for the next month for each shop and item.

- **Source –Kaggle** - Predict Future Sales https://www.kaggle.com/c/competitive-data-science-predict-future-sales
- **Size of Training Data:** 2935849 rows * 6 columns
- **Size of Test Data:** 214200 rows

DATA ANALYSIS

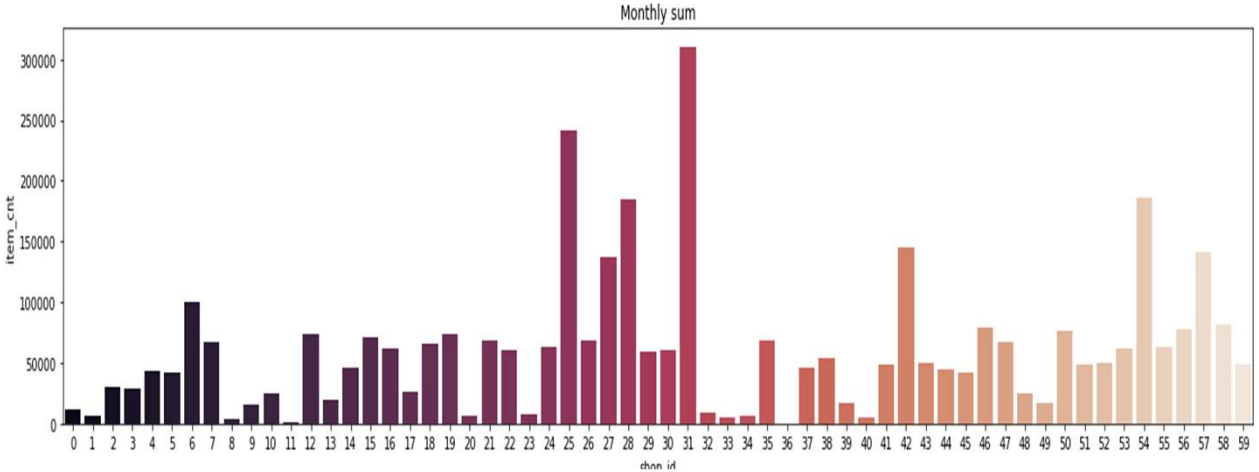


An increase of item sales towards the end and start of the year



Few of the item categories hold most of the sales count.

DATA ANALYSIS

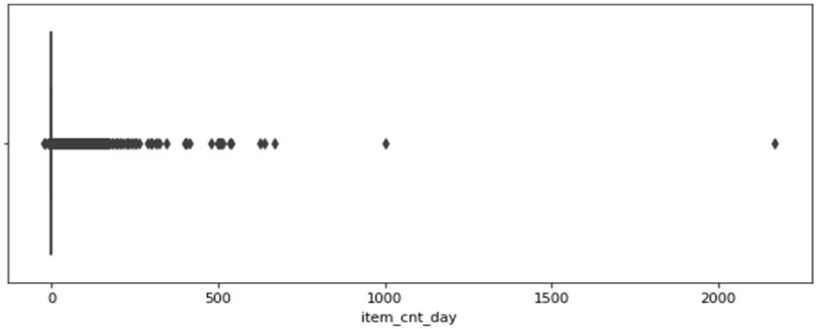


Few of the shops hold most of the item sales count.

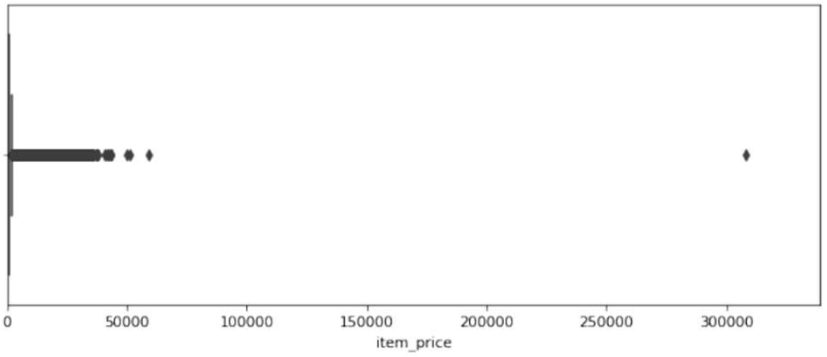
DATA CLEANING

- Checking for NULL Data
- Handling Duplicates
- Checking and Removing Outliers

Plot: Item count per day



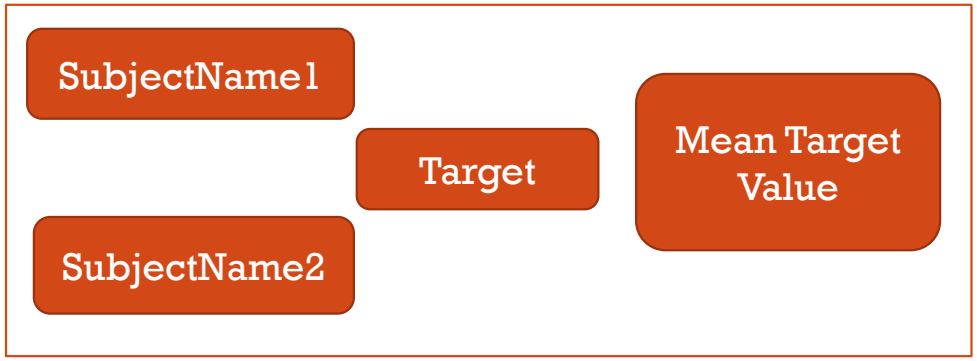
Plot: Item price



FEATURE ENGINEERING

MEAN ENCODING

- Create new features to get more information from existing data.
- Dataframe.groupby(['SubjectName1','SubjectName2'])['Target'].mean()



LABEL ENCODER

- Convert Categorical Data into Numeric Data

LAG FEATURES

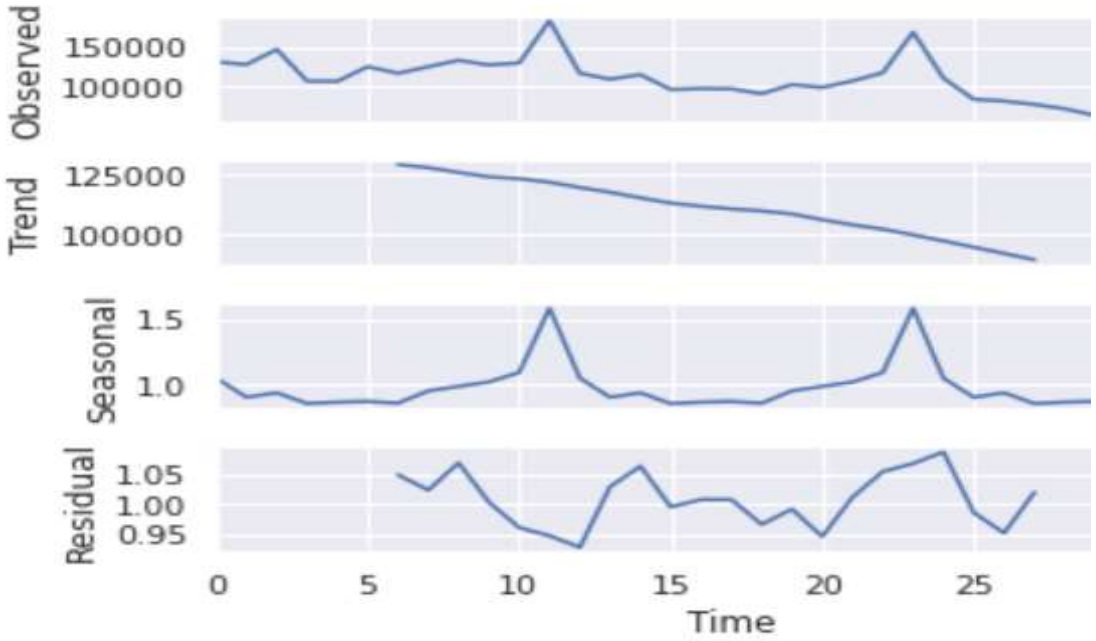
- The value at time t is affected by the value at time t-1.
- Classic way to convert Time Series to supervised learning.
- The past values are known as lags, so t-1 is lag 1, t-2 is lag 2 etc.
- Lags 1,2,3,6,12 to help study the monthly, bi-monthly, quarterly, half-yearly and yearly trends across the year.

Considering Lag = 1, 2 for item count column :

Month	Item count	item_count_lag_1	Item_count_lag_2
Jan 2013	10	NaN	NaN
Feb 2013	20	10	NaN
March 2013	5	20	10
April 2013	2	5	20

TRENDS

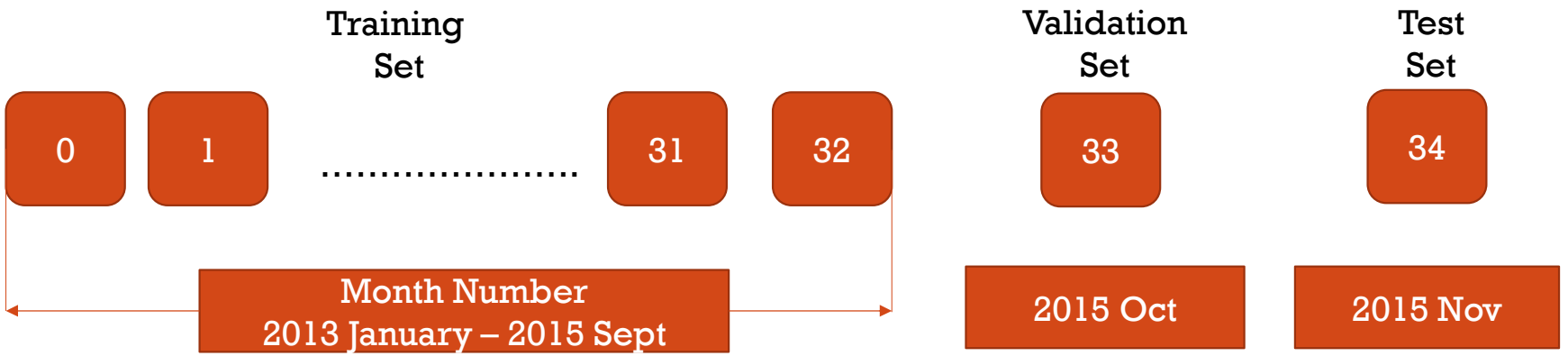
- Created a new feature for the item sales count trends



METHODOLOGY

- Combined the data sources.
- Data Preprocessing and Feature Engineering.
- Train different models and check test results.

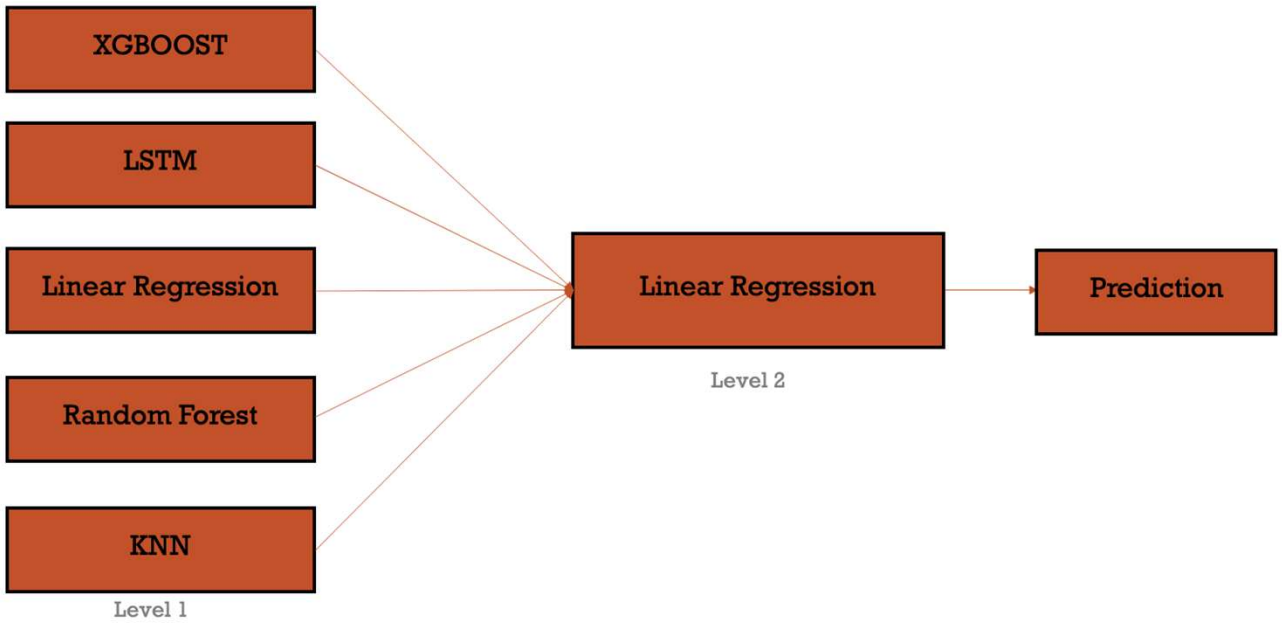
Leave-One-Out Validation Technique



IMPLEMENTATION AND RESULTS

MODELS

- Linear Regression
- KNN
- Random Forest
- LSTM
- XGBoost
- Ensemble model



EVALUATION METRIC

RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Where, yj = predicted value
^yj = actual value
n = no. of data points.

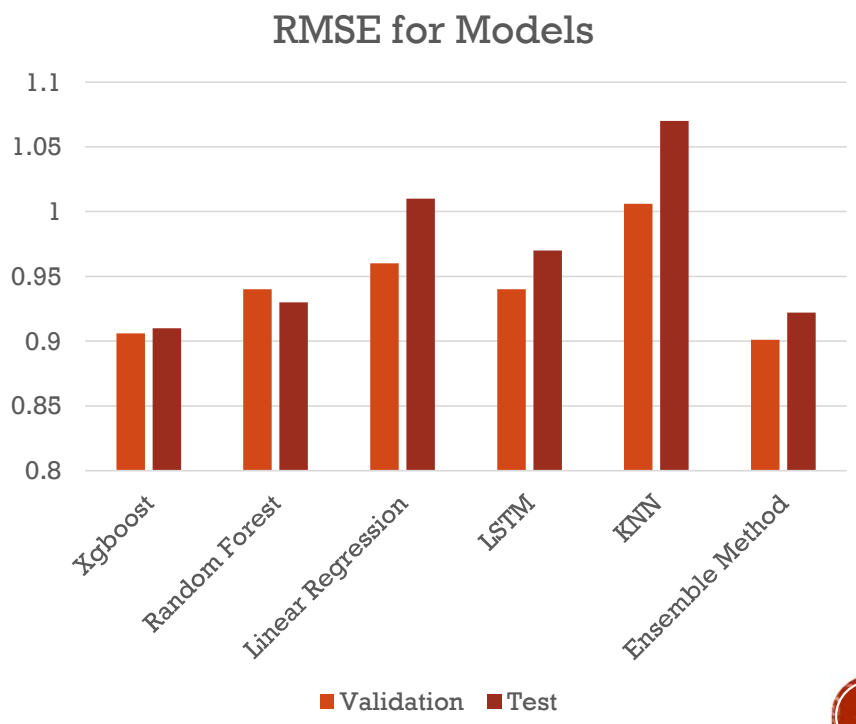
MAE

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Where, yj = predicted value
^yj = actual value
n = no. of data points.

RESULTS AND GRAPHS

Model	RMSE	MAE
Xgboost	Training - 0.82488 Validation - 0.90616 Test – 0.91484	Training - 0.31162 Validation – 0.31585
Random Forest	Training - 0.87622 Validation – 0.94125 Test – 0.93986	Training - 0.32419 Validation – 0.31575
Linear Regression	Training - 0.91687 Validation – 00.96358 Test – 1.01079	Training - 0.34837 Validation – 0.36626
LSTM	Training - 0.88341 Validation - 0.94164 Test - 0.97925	Training - 0.30683 Validation - 0.30250
KNN	Training - 0.6875 Validation – 1.006 Test – 1.07925	Training -0.256 Validation – 0.3267
Ensemble Method	Validation - 0.90180 Test - 0.92209	Validation- 0.31306



CHALLENGES

- Understanding time series data and finding ways to get better prediction result was challenging
- Data was huge, So training models took lot of time.

CONCLUSION

- For prediction, Time Series Components are significant
- The order of data needs to be preserved
- The best forming model is XGBoost

THANK YOU!!