Predict Future Sales

CS580L-01 Spring 2020 Project Final Report
Aditya Chaudhari [achaud12@binghamton.edu, B#00800200],
Darshan Desai[ddesai9@binghamton.edu, B#00816526],
Sahil Mirchandani[smircha1@binghamton.edu, B#00816898].

1. Introduction

- In today's world, small or medium business are struggling to keep their inventory up and successfully predict the sale of product over the year. This not only ensures higher profits but also makes sure that the products stay fresh or updated. Imagine a supermarket selling every brand of a particular products, but end up selling one a few selected and well-known brands. The prediction analysis based on the previous data helps such supermarkets stock products for brands which are more sold than showcasing rest of the brands. They also use prediction to put offers on products, which in turn maximizes the profits.
- In this project we would use machine learning algorithms to precisely predict the sales of the products based on every month of the year. This would help small businesses to cope up with the supply and demand and in turn increase revenue.

2. Problem Statement

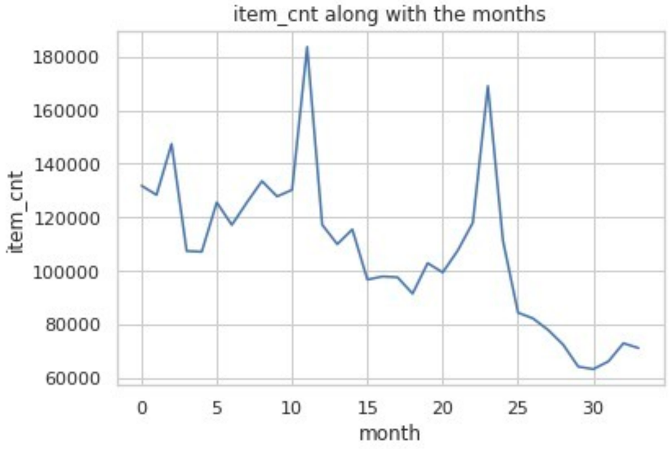
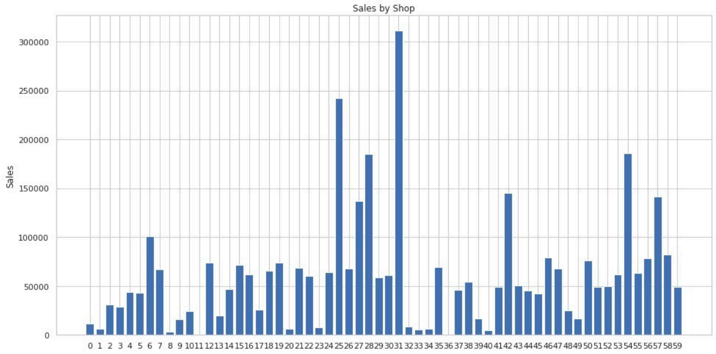
- We are provided with daily historical sales data for one of the largest Russian software firms – 1C Company. The task is to forecast the total amount of products sold in every shop for the test set. The list of shops and products slightly changes every month. We are creating a robust model that can handle such situations is part of the challenge. We are asked to predict total sales for every product and store in the next month.
- We will be predicting future sales based on the above problem statement and dataset obtained from Kaggle [https://www.kaggle.com/c/competitive-data-science-predict-future-sales/]

3. Data Acquisition

- The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge
- The dataset contains data for 60 shops and 22,170 unique items and their sales at each month over the period of 33 months.
- The data is distributed between the 5 files (.csv)
- File descriptions:
    - *sales_train.csv* - the training set. Daily historical data from January 2013 to October 2015.
    - *test.csv* - the test set. You need to forecast the sales for these shops and products for November 2015.
    - *items.csv* - supplemental information about the items/products.
    - *item_categories.csv* - supplemental information about the items categories.
    - *shops.csv*- supplemental information about the shops.
- Data fields

  - *ID* - an Id that represents a (Shop, Item) tuple within the test set
  - *shop_id* - unique identifier of a shop
  - *item_id* - unique identifier of a product
  - *item_category_id* - unique identifier of item category
  - *item_cnt_day* - number of products sold. You are predicting a monthly amount of this measure
  - *item_price* - current price of an item
  - *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
  - *item_name* - name of item
  - *shop_name* - name of shop
  - *item_category_name* - name of item category
- The data was originally in Russian, so it was necessary to convert the data into English.
- They together have sales of 2,14,200 products over the course of 33 months.

4. Data Analysis & Pre-processing

	The total sales of 2,14,200 products which include overall 22,170 unique items sold over the period of 33 months. The graphs shows the relation between the month of the year and the sales every month. We can observe some interesting pattern with the analysis.
	The total sales of 2,14,200 products over 60 small shop owners. The graphs shows the relation between the sales and the shops. From this graph we can observe that which shop has the maximum sales over the course of 33 months

- Data Pre-processing is the crucial part for better prediction model as some points could affect the prediction and can lead to incorrect prediction
- Missing values in the dataset: One can use predictive or averaging techniques in order to fill the missing values. Since our dataset is large and we have many samples, we have removed the missing values. The percentage of missing values is less. Thus, ignoring the

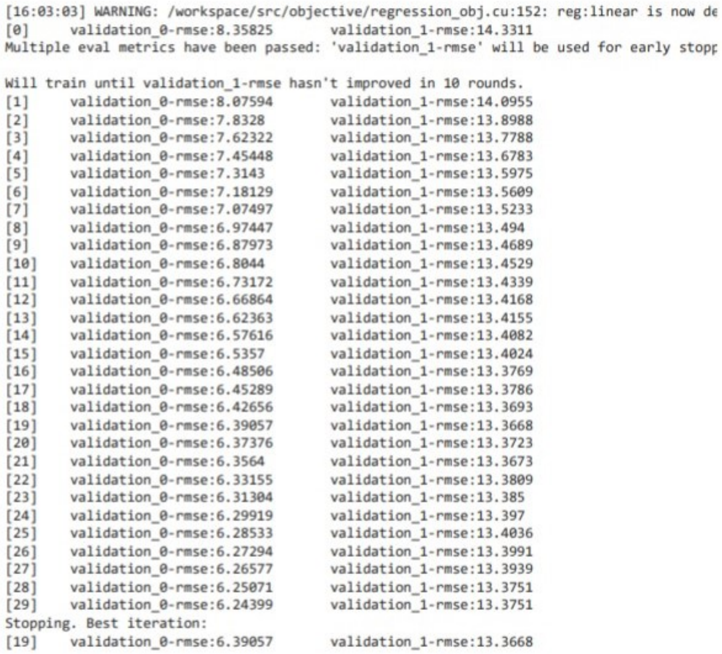
missing values does not create biased estimate in the data analysis. Complete removal of data with missing values results in robust and highly accurate model.

- Duplicates values: Similar to missing values, we need to remove duplicate values. Duplicate values should shift our model in the favor of most duplicated values.
- Outliers/inconsistent data: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. For example, our data contained negative values for item count and item price.
- Standardization: Data standardization is the process of rescaling data for standard prediction.
- Handle same owners of business: We have to perform some specific pre-processing methods based on the dataset. We found that there were different shops/owners with same name which could harm our model. Hence, we need to equate them to one.

After cleaning the dataset: *Number of shops is 42 & number of unique items is 5100.*

5. Model and Results

- We have used two prediction models to predict the sales over the given data – Linear Regression and Gradient Decent techniques. Both the outputs can be compared below and we can notice that the output for Linear Regression is more perfect as the RMSE values of the Linear Regression are more closer.
- Output for Gradient descent, we observe that the RMSE value after 19th iteration is 13.3668 (the best iteration)



```
[16:03:03] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now de
[0]     validation_0-rmse:8.35825       validation_1-rmse:14.3311
Multiple eval metrics have been passed: 'validation_1-rmse' will be used for early stopp

Will train until validation_1-rmse hasn't improved in 10 rounds.
[1]     validation_0-rmse:8.07594       validation_1-rmse:14.0955
[2]     validation_0-rmse:7.8328        validation_1-rmse:13.8988
[3]     validation_0-rmse:7.62322       validation_1-rmse:13.7788
[4]     validation_0-rmse:7.45448       validation_1-rmse:13.6783
[5]     validation_0-rmse:7.3143        validation_1-rmse:13.5975
[6]     validation_0-rmse:7.18129       validation_1-rmse:13.5609
[7]     validation_0-rmse:7.07497       validation_1-rmse:13.5233
[8]     validation_0-rmse:6.97447       validation_1-rmse:13.494
[9]     validation_0-rmse:6.87973       validation_1-rmse:13.4689
[10]    validation_0-rmse:6.8044        validation_1-rmse:13.4529
[11]    validation_0-rmse:6.73172       validation_1-rmse:13.4339
[12]    validation_0-rmse:6.66864       validation_1-rmse:13.4168
[13]    validation_0-rmse:6.62363       validation_1-rmse:13.4155
[14]    validation_0-rmse:6.57616       validation_1-rmse:13.4082
[15]    validation_0-rmse:6.5357        validation_1-rmse:13.4024
[16]    validation_0-rmse:6.48506       validation_1-rmse:13.3769
[17]    validation_0-rmse:6.45289       validation_1-rmse:13.3786
[18]    validation_0-rmse:6.42656       validation_1-rmse:13.3693
[19]    validation_0-rmse:6.39057       validation_1-rmse:13.3668
[20]    validation_0-rmse:6.37376       validation_1-rmse:13.3723
[21]    validation_0-rmse:6.3564        validation_1-rmse:13.3673
[22]    validation_0-rmse:6.33155       validation_1-rmse:13.3809
[23]    validation_0-rmse:6.31304       validation_1-rmse:13.385
[24]    validation_0-rmse:6.29919       validation_1-rmse:13.397
[25]    validation_0-rmse:6.28533       validation_1-rmse:13.4036
[26]    validation_0-rmse:6.27294       validation_1-rmse:13.3991
[27]    validation_0-rmse:6.26577       validation_1-rmse:13.3939
[28]    validation_0-rmse:6.25071       validation_1-rmse:13.3751
[29]    validation_0-rmse:6.24399       validation_1-rmse:13.3751
Stopping. Best iteration:
[19]    validation_0-rmse:6.39057       validation_1-rmse:13.3668
```

Figure 1: Output for Linear Regression

- Output for Linear Regression, we observe that the RMSE value is more better for the test data given along with the problem statement. Here RMSE is 13.95 for test set and 7.35 for the train dataset.
- Linear regression model is a linear model in which we try to fit a line over the training data points.

- We calculate error by using RMSE (**R**oot **M**ean **S**quare **E**rror) value. Which tries to minimize the error by selecting the best values of m and x, in the line equation **y = mx + c.**
- We have performed analysis on our test dataset and which make us predict the item sales on upcoming month of November.
- We found that some of the products have high sales and some have low.
- The box plot analysis shows that the prediction of items in month of November

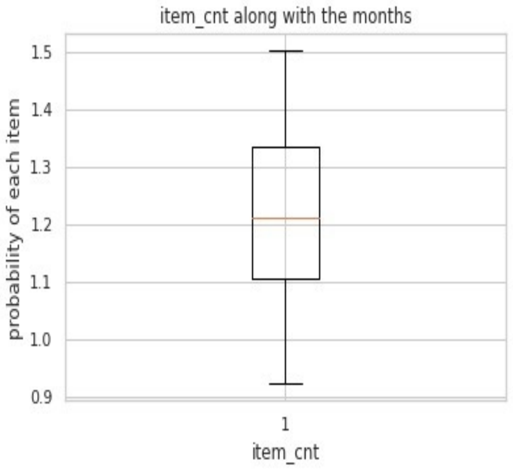
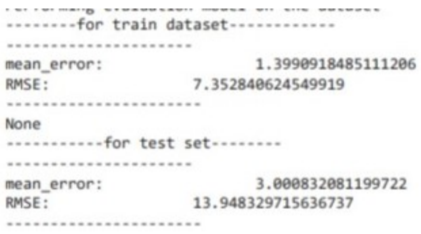


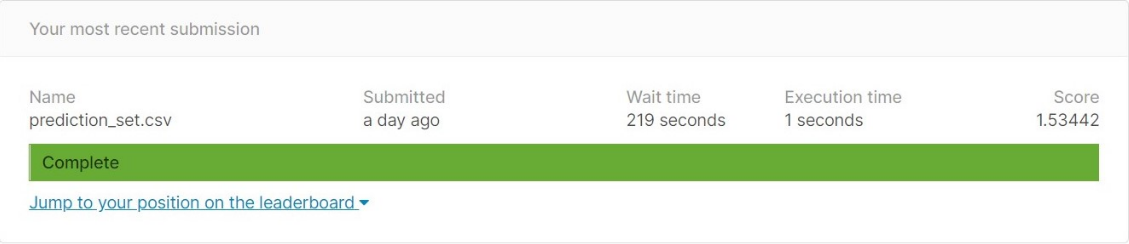
Figure 2: Box Plot, Linear Regression



```
Performing evaluation models on the dataset
--------for train dataset------------
--------------------
mean_error:                 1.3990918485111206
RMSE:               7.352840624549919
----------------------
None
-----------for test set--------
--------------------
mean_error:                 3.000832081199722
RMSE:               13.948329715636737
----------------------
```

Figure 3: Output for Linear Regression

- The results were then submitted in a separate file *'sample_submission.csv'* on Kaggle competition which is currently active.



Your most recent submission

Name	Submitted	Wait time	Execution time	Score
prediction_set.csv	a day ago	219 seconds	1 seconds	1.53442

Complete

Figure 4: Submission on Kaggle.com

6. Project Outcomes

- We have learned that how much Pre-processing takes part while predicting model and how some outliers and inconsistent values can overall result in bad prediction.
- Analysis can picture various insights in the model that can help in verification of our prediction model.
- How to perform linear regression and gradient decent algorithm.
- How to identify best algorithm that fits our data model.
- How to choose best algorithm from the pool of machine learning algorithms
- Different models could be used to verify our prediction model.
- Neural network-based algorithms are slower.