Report On Future Sales

By-Saran Tejas Kaja

December 13, 2019

Table of Contents

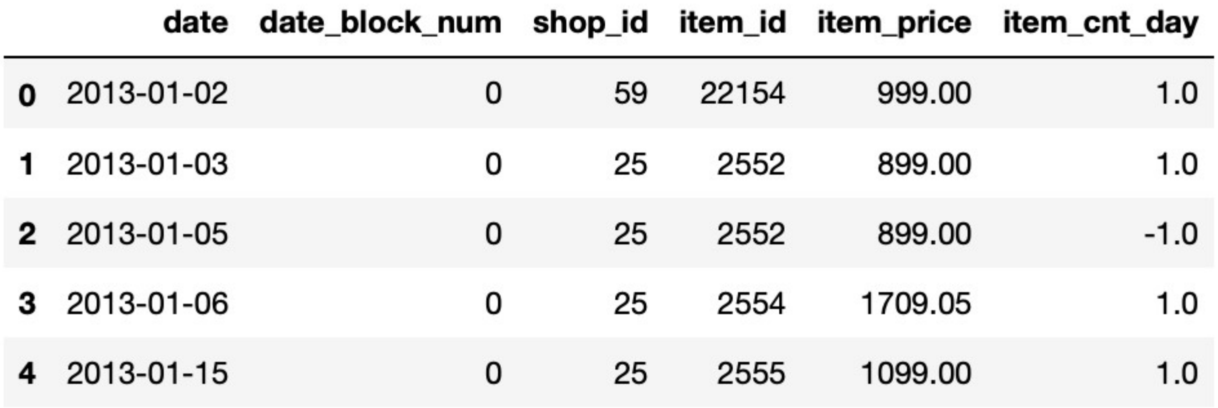
Summary

The project consists of analysis performed on the dataset provided by a Russian firm 1C Company. It has been obtained from Kaggle competition to Predict Future Sales. The dataset consists of about 3 million records and 6 attributes which records the sales of various products in different stores on an everyday basis. We try various classification algorithms such as KNN, Naïve Bayes and Decision tree to predict the sales for a given day. Next, we try to implement various timeseries forecasting using ARIMA, SARIMA and LSTM. The results have been compared at the end to draw the conclusions. This project has been implemented using various python libraries.

Introduction

The data has been obtained from Kaggle, from an active competition Predict Future Sales which is provided by one of the Russian software companies 1C Company. The company specializes in software development, publishing and support. The dataset consists of the historical daily sales data from January 2013 to October 2015, that tracks information such as item, its price, the quantity sold and the shop where it has been sold from.

The dataset obtained from Kaggle consists of six csv (comma separated files) including a test file and a sample submission file for the competition. The other files include items, item_categories, shops and sales_train. Here, the items and item_categories records each item for the company and shops records the details of the shop. All of the information is then combined in the sales_train file. For the purpose of this project we are only using sales_train file. The sample data for the sales_train file is as shown in the figure below.



	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	2013-01-02	0	59	22154	999.00	1.0
1	2013-01-03	0	25	2552	899.00	1.0
2	2013-01-05	0	25	2552	899.00	-1.0
3	2013-01-06	0	25	2554	1709.05	1.0
4	2013-01-15	0	25	2555	1099.00	1.0

The dataset consists of 6 attributes viz., date, date_block_num, shop_id, item_id, item_price and item_cnt_day. Date is the date of transaction starting from January 2013 to October

2015; date_block_num is the consecutive number assigned to each month such as January 2013 as 0, February 2013 as 1 and so on; shop_id and item_id are the unique identifiers for shop and items respectively; item_price is the price of the item being sold on that day and item_cnt_day is the quantity of the items sold which is also the value being predicted for this project. There are 34 date blocks, i.e 34 months of historical data being considered with about 60 shops and nearly 22,000 products to predict the sales.

For predicting the sales, the data is first divided into training and testing data to train the models on the data. The using the testing data we predict the accuracy for the model to predict the future sales. We fit KNN, Naïve Bayes and Decision tree classification models for prediction. Next, since the data is continuous in time, we fit timeseries forecasting ARIMA, SARIMA and LSTM models for prediction.

The company collects the sales information of its products from multiple shops. It also collects the price and the inventory maintained. The aim of this project is to harness the information collected to be able to predict the sales in the future. The prediction of the future sales would enable the company to meet its sales requirements. It would also help the company to manage its inventory based on the demand and thus reduce the waste created due to excessive number of items. This would also help in resource management for the company to be able to meet the supply and demand. This would help the company in terms of cost along with its impact on environment with lesser waste.

This topic was very interesting since, sales prediction is relevant to most of the industries that we might work for in the future. Most industries, irrespective of the domain, works towards a better inventory and resource management. While at the same time waste reduction will have a lasting effect on the environment; lesser the waste, lesser the harm to the already dying planet.

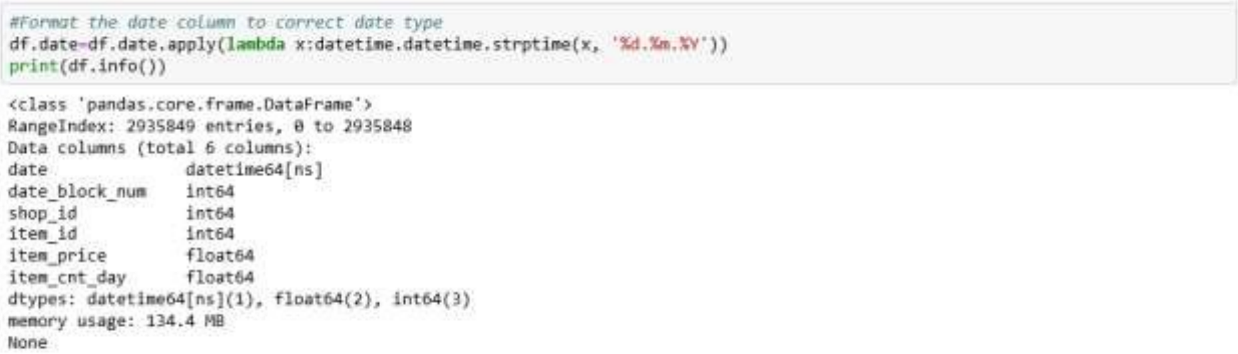
Similar predictions can be utilized for crime prediction to manage the dispatches and staff in a police department or to manage the number of beds in emergency services. It could also be utilized in the food industry to reduce food waste and meet the demand. Working with a large data such as this with about 3 million records would be something that we might be encountering in future on a daily basis making it a good dataset to work on as a part of academic project.

Implementation

The data was downloaded from Kaggle.com which was then imported into the Python Jupyter notebook using pandas libraries. Once, the data was imported next step was to clean the data.

Data Cleaning:

The first step was to check if the datatypes of the imported data was correct. Most of the features were imported with appropriate datatype except the date field which was imported as an object. So, it was converted to the datetime datatype.



```
#Format the date column to correct date type
df.date=df.date.apply(lambda x:datetime.datetime.strptime(x, '%d.%m.%Y'))
print(df.info())
```

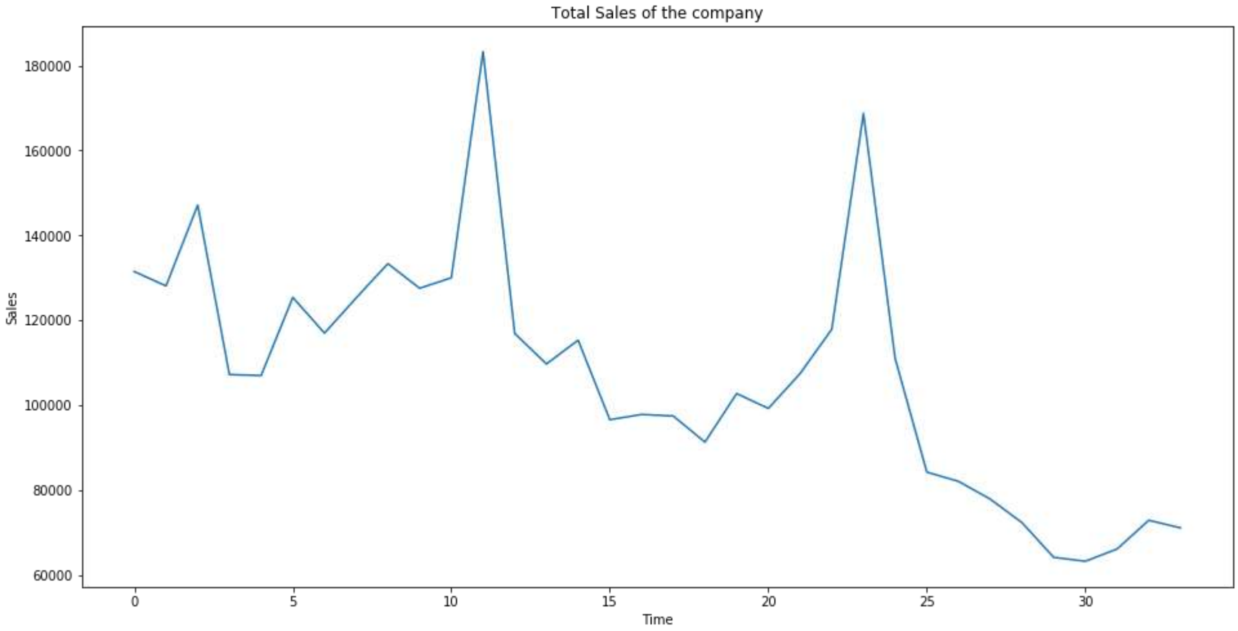
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2935849 entries, 0 to 2935848
Data columns (total 6 columns):
date              datetime64[ns]
date_block_num    int64
shop_id           int64
item_id           int64
item_price        float64
item_cnt_day      float64
dtypes: datetime64[ns](1), float64(2), int64(3)
memory usage: 134.4 MB
None
```

Next, we check if there are any null values in the dataset, but the dataset was comparatively clean with no null values. However, we found about 7300 negative values in the item count column. The values are almost negligible compared the dataset, so we could have eliminated those rows. But since it counts the number of sales of an item, the negative integer could indicate an item returned for any reason. So, assuming the negative values to be returned item, we decided against eliminating the row.

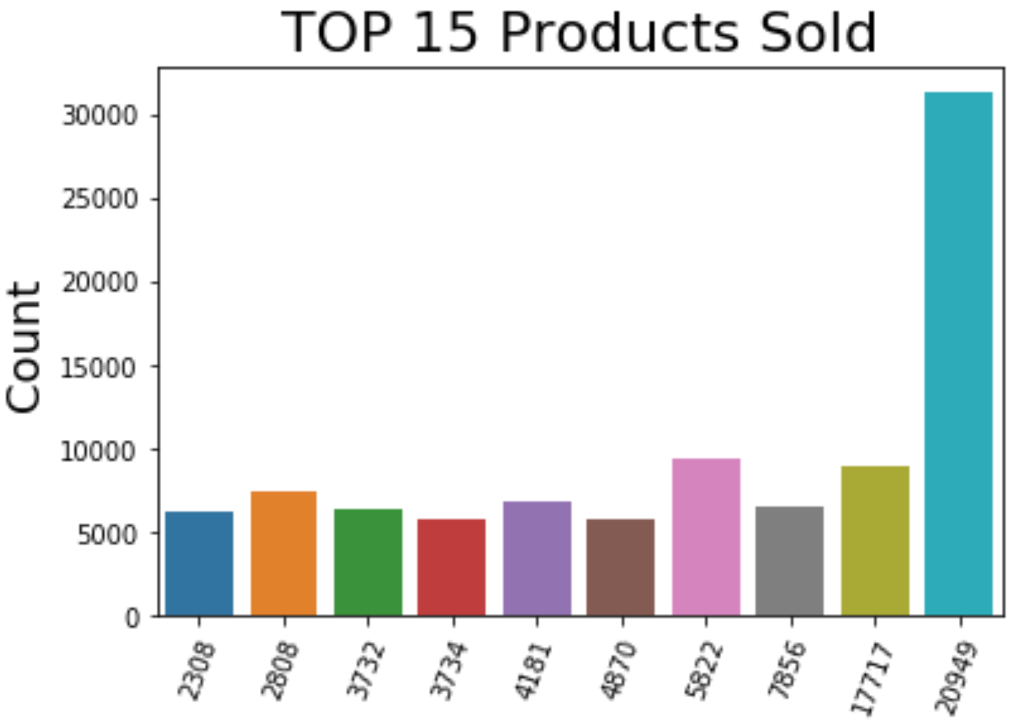
After the data cleaning, we prepared the data to be used to fit in the models. As the we divide the data into training and testing data with 80:20 ratio. For the purpose of prediction, we ignore the date column and hence, it is dropped from the train and test dataframes.

EDA

To understand the data, we plot a line plot for the sales over time. Here, we observed that the sales has reduced over the time. But it can also be observed from the graph that the sales usually peak around the 12 and 24 on time chart i.e. during the month of December. The plot also shows a similar tread of gradually increase towards December then drop post December.



Next, we plot the top 15 products sold. Here we see that product with item id 20949 was their highest selling product. The mean item price was 890.85 while the highest price was 307980.



Now the data was clean, prepared and we had the general idea about the data, next step was to fit the models to for prediction of the item count i.e. item_cnt_day feature. We compared KNN, Naïve Bayes and Decision tree models before the timeseries forecasting models.

KNN

K-nearest neighbor or KNN is one of the simplest algorithms used for the classification of the dataset that classifies the value based on the highest occurrence of values of the k nearest neighbors. Here, k is the assumed number of neighbors that would be considered to predict or classify the new variable to one of the classified groups. The value of k considered is always an odd number to eliminate the probability of equal number of neighbors in any given case. The distance between the neighbors can be calculated by means of Euclidean Distance or Manhattan Distance.

Euclidean Distance, d_e, is calculated as:

$$d_e = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Manhattan Distance, dm, is calculated as

$$d_m = |x_1 - x_2| + |y_1 - y_2|$$

Here,

(x1, y1) = Coordinates of the available value

(x2, y2) = Coordinates of the value to be predicated

Naïve Bayes

Naïve Bayes is a classification algorithm that uses Bayesian Inference based on ‘If A then B’. This provides the posterior probability which is a belief that the hypothesis is true between the values of 0 and 1. This is calculated as product of likelihood and prior when divided by probability of marginal value. This can be defined in mathematical form as:

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)\ P(hypothesis)}{P(data)}$$

Decision Tree

Decision Trees are the simplest forms of the trees constructed to come to a conclusion that is derived from rules such as ‘if.. else..'. Decision trees are split the values into 2 child nodes. This process is repeated until the error doesn’t decrease or to a minimum number of points is reached. Finally, the tree is pruned to a minimum total error.

$$C_\alpha(T) = \sum^{|T|} \quad \sum \quad e_i(c_m) + \alpha|T|$$

$$m=1 \quad x_i\epsilon \; R_m$$

ARIMA

ARIMA or Auto-Regressive integrated moving average, this is one of the famous model which helps in solving time series problems for a long time. This method helps us in solving the problems with trend with ease, we do not have to linearize the trend before implementing this model as this has the capability to work even with the trends.

This model takes the lags of the data and the error bits and differentiates the trend to make it stable based on all these three components, it evaluates the data and it works in a linear fashion. This method uses the data from past events and assigns them weight according to their age, the error is also calculated in the similar manner for the corresponding events and that is used to predict the future events. The problem with this model is it doesn't take seasonality into consideration that is one of the biggest disadvantages.

SARIMA

This is termed as Seasonal Auto-Regressive Integral Moving Average (SARIMA), this model is an extension or upgraded version of ARIMA model. This algorithm takes the seasonality component of a timeseries into consideration and predicts the future events, this works in a similar way to ARIMA model and the only addition is it takes seasonality into consideration due to which this is better than ARIMA if we have seasonality in our data.

LSTM

Long short term memory in short known as LSTM is one of the famous algorithm used in neural networks, nowadays this algorithm is gaining lot of transaction and is most widely used in businesses to forecast the trends and quantities, it is mostly used to forecast sales, costs, lifetime of a product etc. It has many advantages over the general time series models such as ARIMA and SARIMA, the biggest advantage is that this model learns from the data without linearly going through the previous data, because of which it needs a lot of data and time to learn from the data, once it does it, it aces the race and predicts better than any other existing models. But the drawback this model has is it needs lot of data and time to train the model.

It takes three inputs which has the present data the previous and the future data. it stores the historical data and joins with the present data to compute the output. It uses sigmoid and tanh functions through its way to get rid of error bits and learns about the data by comparing the present state with future and past states. The output of this model can also be modified by the activation function we are using in the final stage.

Data Analysis

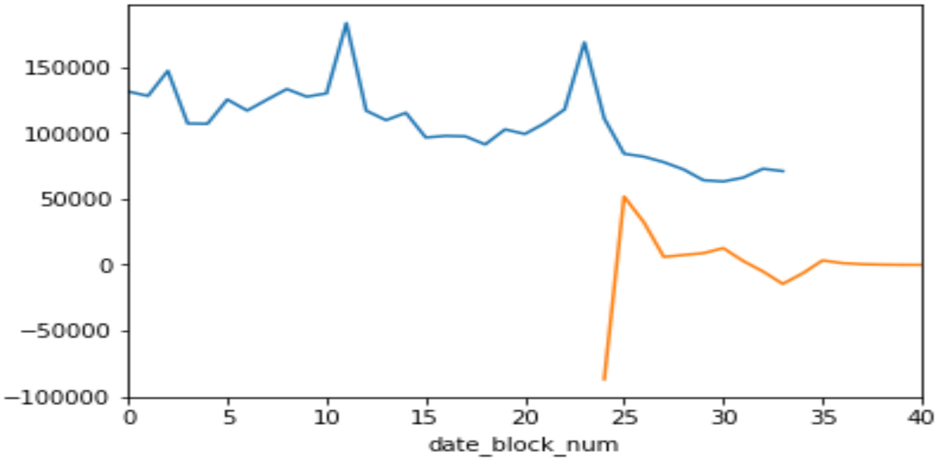
Once the classification models were fit for the training data. We obtained the accuracies for the comparison. The comparison can be seen from the table below.

Algorithm	**Accuracy Score**
K-Nearest Neighbors	89.61%
Naïve Bayes	88.81%
Decision Tree	87.54%

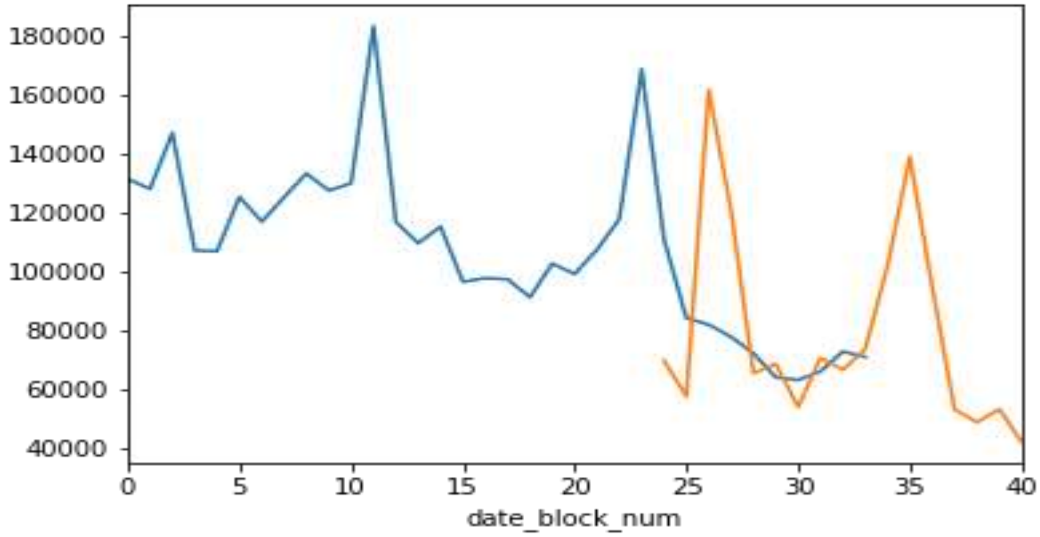
Table 1. Comparison of the accuracy scores

From the table it can be seen that all the classification algorithms give about the same rate of accuracy. KNN has the highest accuracy while Decision tree has the lowest for this dataset.

The first timeseries model that we had considered was ARIMA. Here, the values of autoregression component (p), integrated component (d), and moving average (q) are considered 1, 2 and 3 initially which was then changed to 1,2 and 2 respectively to have the least AIC. This returns the plot as follows.



But as it can be seen from the plot, the predicted values are further from the actual values. Hence, we implemented the SARIMA model with an additional factor of seasonality. Since this is a yearly sales data, there will be a seasonal factor present in the data which can be seen from the figure 1 in EDA. The factors in SARIMA models are seasonal autoregression component (P), seasonal integrated component (D), seasonal moving average (Q) and time period (M) which is taken as 1, 1, 0 and 12 respectively. Since the time period is yearly, the value of M is 12. This returns the following plot. It can be seen that the plot is nearly close to the actual data and also able to predict the seasonal hike in the sales.



Next we implemented the LSTM, which uses neural networks for time series. Here, with 5 epochs the mean squared error obtained was 7.13.

Discussion

This project gave us a scope to deal with time series data and the models that can be used to do predictions with such data. It was interesting to learn about these methods and models as this is the most required skill and data a data analyst would handle with in any industry as every industry has data with respect to time. We also got a chance to learn about some algorithms which we never used in class such as ARIMA, SARIMA and LSTM, they were vast and the algorithms which we learned in the class helped us to learn about these algorithms with ease as we had the knowledge to interpret them. We liked the LSTM algorithm the most as it was related to the neural network family and was very powerful. It is widely used in the industry nowadays to solve most of the problems related to time series as it selects the data non linearly and learns from them to do the predictions due to which it will yield better results in the long run. In the next six months we plan to learn more about LSTM and try to run the whole model using more computation power and the learn various ways to tune it to optimize the results.

References

Predict Future Sales. (n.d.). Retrieved from https://www.kaggle.com/c/competitive-datascience-predict-future-sales

What is the Advantage of using RNN and LSTM over traditional methods for time series of streaming data? Retrieved from https://www.researchgate.net/post/What_is_the_advantage_of_using_RNN_and_LSTM_over_traditional_methods_for_time_series_of_streaming_data