Future Sales Prediction

Pallava Arasu Pari

Madhu Shri Rajagopalan

Deepti Gupta

Charannag Devarapalli

Yash Anilbhai Shah

Harrisburg University of Science and Technology

Predict Future Sales

Introduction

In this competitive world, where the number of businesses built are increasing at a quicker rate, companies on any scale will thrive towards increasing its sales and find ways to stock products to minimize costs and maximize profits. Forecasting finds its application in almost any industry, a few examples include predicting stock in stock market, predicting scores of a football match and like. Companies are leaning towards using historical data to find patterns to predict sales and make informed decisions on managing inventory with proper product stocking plans. Data Science plays a major role in achieving this objective using machine learning techniques to forecast sales.

In this project, a similar situation dealt by a large Russian software firm, 1C company which is trying to predict total sales for every product and store in the next month is been tried to help with by using Machine Learning techniques. The company has provided a challenging time-series dataset consisting of daily sales data. This work is focused on utilizing a variety of models to predict sales. As a baseline method, regression decision tree is implemented. Then, a more sophisticated model of grouping sales data individually for each shop and then performing linear regression on the individual shop model was implemented.

Related Work

This method of time series forecasting is popularly implemented with statistical approach using ARIMA and by using machine learning method of Artificial Neural Networks. These two approaches have been proven successful in time series sales forecasting and stock predictions. In addition to this generalized technique, an interesting work of a co-competitor for the analysis on

the same dataset utilized a step by step approach to solve this problem. This work included Feature engineering and scaling and then a variety of modelling namely, Linear regression, Light GBM, Neural Network, finally an ensemble technique for boosting the performance.

Data Overview

The dataset contains daily historical sales data of over 22,000 items from 60 different shops for a date range from January 2013 to October 2015. The major attributes in the training data include date, date_block_num, shop_id, item_id, item_price, item_cnt_day. The training and testing data are provided by the kaggle competition.

date	date in format dd/mm/yyyy
date_block_num	consecutive month number
shop_id	unique identifier of a shop
item_id	unique identifier of a product
item_price	current price of an item
item_cnt_day	number of products sold

Exploratory Data Analysis

Exploratory data analysis not only provides useful insights into the data, but it lets you avoid using inaccurate models on the data. Under this part, we found out that data did neither have any missing values nor outliers. Though at the first glance, we found one outlier, but concluded it as a genuine sale after further research.

A tableau report was also created under this phase to dig more about the potential trends and patterns in the data. This report uses the data aggregated for the sale of items month wise where month is the continuous month number (0-33) for 3 years.

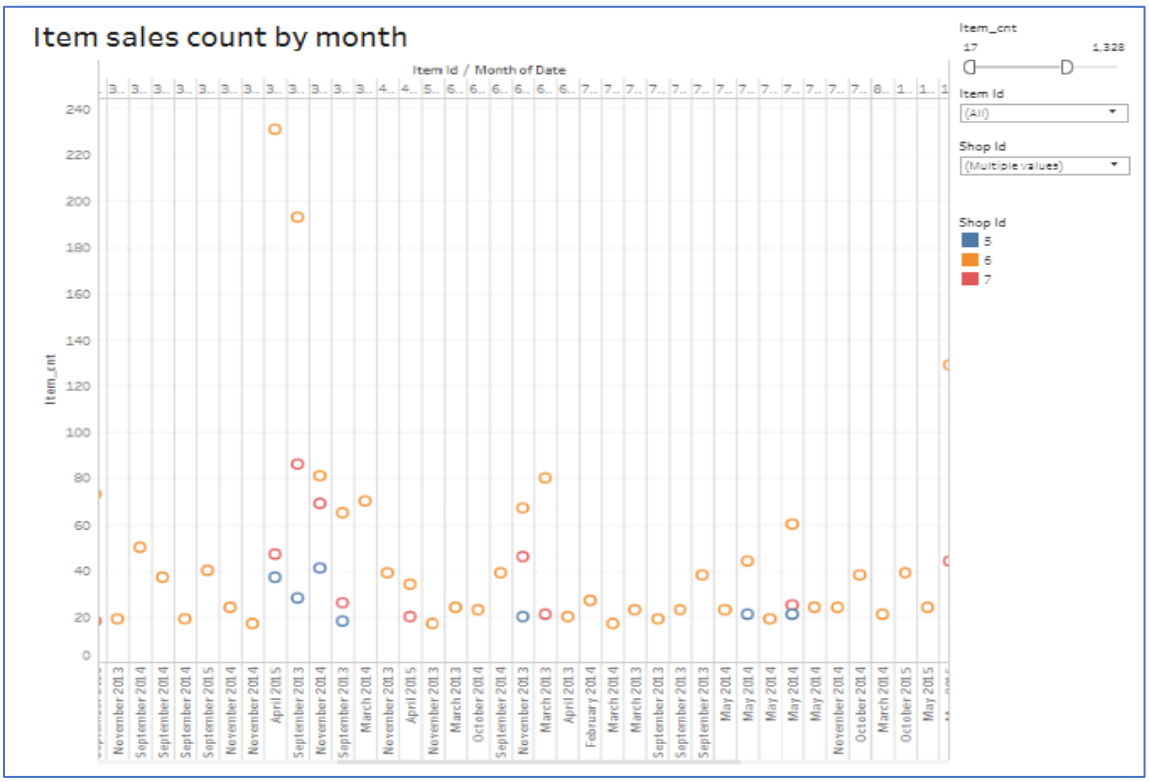


Figure 1. Tableau Report to visualize item count by month (Click here to navigate to report)

Technical Approach

Method 1: Decision Tree Classification

To predict the sales for a data where the model can easily learn, supervised learning method - decision tree was our first choice.

The data was grouped by items sold in each shop and for each item in the shop. Using library “sklearn”, a decision tree model was trained to predict the item count sold for each shop and item provided in the testing data. Inserted below is the file exported containing the predicted

records. We tried forming a decision tree by splitting shop ids on the root level and gradually advancing to item ids and date block numbers, but it remains a challenge so far.


SalesTest_DecisionTreeClass.csv

Method 2: Decision Tree Regression

The data prepared for the decision tree classification was used to implement regression on trees. Using library "sklearn", a decision tree regressor was trained to predict the item count sold for each shop and item provided in the testing data. Below is the file exported from the predicted results from decision tree regression.


SalesTest_DecisionTreeRegr.csv

Method 3: Designing Linear Regression Model

We successfully predicted the results from the two algorithms and still knew that our model was not strong enough to be trained to relate to every shop_id, item_id and date_block_number combination.

We worked on our next approach where we created a model for every shop_itemid sequence but since this is a distinctive design, a prototype was formulized only for shop id = 5 and shortening our dataset from 2,935,849 rows to 38,179, considering the size of the data, training time and memory of our computers.

A combination of a shop and items sold in that shop is separately treated as a model with the dictionary key = "Shopid_itemid". That is, for a single row of input data that contains a shop id and an item id, is predicted using a model that is specifically developed for that shop and that particular item.

The data was aggregated in dictionaries to get the total count of item sold in that month and was inserted with item count as 0 for those months in which no sales were made in order to train the model accurately for the months with sales and no sales. A shop might not have sold all the items constantly on all the months. Below is a typical dictionary with key = 5_1830 for which we devised a linear model.

Shop_5 - Item_1830
5_1830

	date_block_num	item_id	item_cnt_agg
0	0	1830	0.0
1	1	1830	10.0
2	2	1830	16.0
3	3	1830	14.0
4	4	1830	7.0
5	5	1830	5.0
6	6	1830	11.0
7	7	1830	16.0
8	8	1830	11.0
9	9	1830	4.0
10	10	1830	4.0
11	11	1830	2.0
12	12	1830	4.0
13	13	1830	4.0
14	14	1830	3.0
15	15	1830	5.0
16	16	1830	2.0
17	17	1830	2.0
18	18	1830	1.0
19	19	1830	2.0
20	20	1830	2.0
21	21	1830	1.0
22	22	1830	1.0
23	23	1830	2.0
24	24	1830	2.0
25	25	1830	1.0
26	26	1830	1.0
27	27	1830	0.0
28	28	1830	1.0
29	29	1830	0.0
30	30	1830	0.0
31	31	1830	0.0
32	32	1830	0.0
33	33	1830	0.0

Using library “statsmodel”, a linear regression was performed on the prepared data. A low R-square can be explained from the large number of months that contain zero sales. The normality conditions for the model were also verified. An example is shown below to display the trained model for shop id 5 and item id 1830.

5_1830

OLS Regression Results

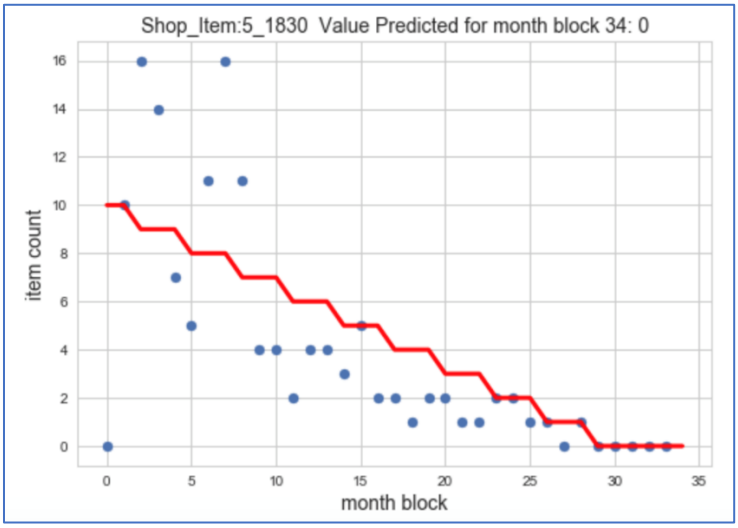
Dep. Variable:	item_cnt_agg	R-squared:	0.522
Model:	OLS	Adj. R-squared:	0.507
Method:	Least Squares	F-statistic:	34.94
Date:	Fri, 12 Apr 2019	Prob (F-statistic):	1.41e-06
Time:	12:50:19	Log-Likelihood:	-87.782
No. Observations:	34	AIC:	179.6
Df Residuals:	32	BIC:	182.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.5630	1.107	8.642	0.000	7.309	11.817
date_block_num	-0.3407	0.058	-5.911	0.000	-0.458	-0.223

Omnibus:	6.338	Durbin-Watson:	1.040
Prob(Omnibus):	0.042	Jarque-Bera (JB):	7.731
Skew:	0.277	Prob(JB):	0.0210
Kurtosis:	5.269	Cond. No.	37.6

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

There were few shops and items ids which were present in testing data but not in training data for which our model predicted NA. Below is an example of our prediction where the model 5_1830 predicted sales for the next month, i.e. date block number 34. It has predicted 0 sales looking at the past trends of the sales. Blue dots are the actual sales and red line is the predicted values from our model.



Inserted below is the file exported containing the predicted records from linear regression model.


SalesTest_LinearRegression.csv

Conclusion

With this project, we have designed three approaches to predict future sales, first two with decision tree classification and decision tree regression showing an example of supervised learning method. To devise a better approach for a strong model training mechanism, we moved to the third and final approach of implementing “linear regression” by creating a regression model for every shop_item combination.

This approach is better than others because it formulates a strong relation between item and consecutive months numbers for a given shop. As a part of deliverables, we also generated a csv file with the predicted data for the given shop and items in the test dataset.

References

Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction," Journal of Applied Mathematics, vol. 2014, Article ID 614342, 7 pages, 2014. doi:10.1155/2014/614342

Mashael A. Al-Barrak and Muna Al-Razgan (2016, July), "Predicting Students Final GPA Using Decision Trees: A Case Study" Vol 6. General Format Retrieved from http://www.ijiet.org/vol6/745-IT205.pdf

Asher B. Curtis, Russell J. Lundholm and Sarah E. Mcvay (Summer 2014), "Forecasting Sales: A Model and Some Evidence from the Retail Industry" CAR Vol 31. General format retrieved from http://faculty.washington.edu/smcvay/CLM_Final.pdf