Day Based Sales Amount Forecasting of Dairy Products

Mert Kelkit, Rümeysa Eliöz and Taha Bilal Özbey
Department of Computer Engineering
University of Marmara
Göztepe, Kadıköy, İstanbul, 34722, Turkey
{mertkelkit & rumeysaelioz & tbilal}@marun.edu.tr

Merve Dereboylu and Nazım Berke Demir
Department of Industrial Engineering
University of Marmara
Göztepe, Kadıköy, İstanbul, 34722, Turkey
{mervedereboylu & nazim.berke}@marun.edu.tr

***Abstract* - This paper presents several data science methods, which are K-means clustering Algorithm, Linear Regression, Polynomial Regression, Stochastic Gradient Descent Regressor, Linear Support Vector Regression. For this study, the data used is cold product sales of a grocery store chain.**

I. Introduction

It is known that estimating future is one of the most important powers of an organization. Because, future strategies are based on this estimations. Sales amount forecast has great importance for organizations forwhy sales amount affects production processes, row material usage, labor force usage etc. This resources cause a lot of cost to organization.

Machine learning is a technique which uses mathematical and statistical methods to deduce results from existing data and makes estimation by using these results. Steps, that take part in machine learning process, are data mining and prediction modeling. Both of these steps require data search to look for patterns and adjust program actions.

This paper presents studies which contains the 3-year data set of dairy product sales in a market chain. This data was used to estimate using several machine learning methods which are K-means clustering Algorithm, Linear Regression, Polynomial Regression, Stochastic Gradient Descent Regression, Linear Support Vector Regression with aim to show which method is more efficient.

II. Related Work

Sales forecasting is one of the most important agendas in business life. However, this is a difficult task for employees to do, and so many software tools have been developed. Three of these are:

II.A. Effect Manager

This tool is one of the most widely used for forecasting sales. Effect Manager has made a unique algorithm to calculate an accurate baseline based on POS data or your own sales in data. The raw baseline from historic data is calculated by deducting promotion sales and seasonality from the total sales. This algorithm calculated the future baseline by looking at the current trend and adding seasonality. This gives an accurate baseline without promotion sales [1].

II.B. SalesLoft

SalesLoft is the sales engagement platform, helping sales organizations to deliver a better sales experience for their customers. More than 2,000 customers use the company's category-leading sales engagement platform [2].

II.C. ZohoCRM

ZohoCRM is a unified platform that helps you juggle multiple business activities efficiently. It's designed to help you sell, market, analyse, manage better and collaborate with customers and employees. You can gather powerful insights from customer interactions through various communication channels like email, phone, social media, and more [3].

III. Approach

As the first part communications between the Danone Dairy company have begun. 3 years of sales data was acquired in order to continue our studies. The names in the data of course was changed (the names are product 1, product 2, brand 1, brand 2...) in order to maintain the privacy of the company. After acquiring the data the first thing was visualization. Different tables and graphics have been made to further understand the data and interpret accordingly.

After, K-Means clustering applied to data in order to understand and interpret the distribution of data. Since target of this study is continuous (total sales amount in given day), regression algorithms are tested with the data.

While predicting the total sales amount on given day, expected input parameters are a day in the future e.g. 15 April, 2021 and estimated change in inflation rate on given date. According to previous data, trained model finds average total sales revenue/unit price of given day e.g. 15 April 2016, 2017 and 2018 and predicts a sales amount according to day, month, year, estimated change in inflation rate, estimated total sales revenue/unit price.

IV. Experiment Setup

IV.A. Clustering Experiments

K-Means algorithm is used for clustering. All features were used for clustering experiments. Since some attributes are nominal, label encoding is applied. For example, Product 0 was replaced with 0 and Brand 1 was replaced with 1 and so on.

K-Means algorithm has the hyper parameter k, so k parameter is changed through experiments within the interval $[4, 15]$. The Sum of Squared Error (SSE) is widely used to determine the optimum number of clusters and evaluation. The clustering with the lowest value of SSE gives the best result.

Product 0, Brand 0, Profile 1, 20th day, 12th month, 2018 is the most occurring values. Average sales revenue is 2045.62, average unit price is 6.420, average inflation rate is +1.101 and average sales amount is 525.34.

IV. B. Regression experiments

For regression experiments, linear regression, polynomial regression, stochastic gradient descent and support vector regression algorithms are tested.

For linear regression experiments, different proportions of train – test sets and different random seeds are used. For polynomial regression experiments, additionally to linear regression parameters, different polynomial degrees are tested. For stochastic gradient descent and support vector regression experiments, additionally to linear regression parameters, different regularization terms, epochs and penalty parameters are tested.

For each experiment with different parameters, mean squared error, mean absolute error and R-squared scores are calculated which are the most reliable regression evaluation metrics. Since mean absolute and mean squared errors can have any positive value and it's better when they are lower; R-squared score have values between 0 and 1, inclusively, and it's better when it's higher, equation (1) is derived and its calculated for each experiment. While evaluating models, main objective is to minimize the result of equation (1) and best model chosen according to result of equation (1).

$$Result = Normalize(Mean\ Squared\ Error) + Normalize(Mean\ Absolute\ Error) + (1 - R^2\ score) \quad (1)$$

V. Experimental Results

V. A. Clustering Experiments Results

Experiments results with around 85K records are below:

# Of Clusters	Avg. # of instances in Clusters	Standard Deviation	SSE	Silhouette Score
4	21371,7	18399,3	1e+11	0,5644
5	17097,4	18150,1	66e+9	0,5631
6	14247,8	15815,2	51e+9	0,5341
7	12212,4	12654,2	42e+9	0,4980
8	10685,8	11047,8	35e+9	0,4796
9	9498,5	10605,7	31e+9	0,4740
10	8548,7	10372,1	28e+9	0,4718
11	7771,5	9055,4	25e+9	0,4448
12	7123,9	7924,6	23e+9	0,4187
13	6575,9	7863,5	21e+9	0,4346
14	6106,2	7755,1	19e+9	0,4343
15	5699,1	6729,7	18e+9	0,4075

TABLE 1 – K-MEANS CLUSTERING METRICS RESULTS

The average numbers of instances in clusters is decreased as number of clusters increases. Standard deviation, sum of squared error (SSE) and silhouette score is as well.

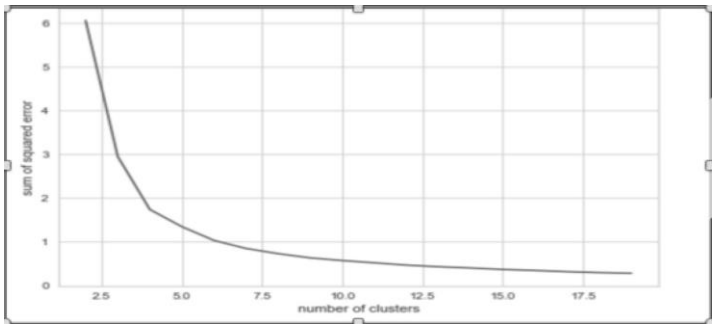

TABLE 2 – SSE VS NUMBER OF CLUSTERS

Elbow method is used to find the optimum number of clusters as table 2. The point at which the graph is tapered gives us the optimal k number so optimum k value is 6.

Attributes	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Product Number	101	92	100	127	114	74
Brand Number	6	4	6	7	7	6
Profile Number	1	1	1	1	1	1
Day	15	16	16	16	16	17
Month	6	6	6	7	7	7
Year	2017	2017	2017	2017	2017	2017
Total Sales Revenue (₺)	6990	632	4270	13239	2120	41359
Unit Price	5,77	7,44	6,6	5,46	4,64	6,11
Inflation	1,25	1,07	1,13	1,27	1,09	1,23
Total Sales Amount	1547	196	902	2569	665	7171

TABLE 3 – CLUSTER CENTROIDS

In table 3, Profile number is not decisive because it is 1 in all clusters, same as day, month and year. However, we see that the total sales amount is different in all clusters and the values between them seem very different. Therefore, it can be said that total sales amount is based on clustering.

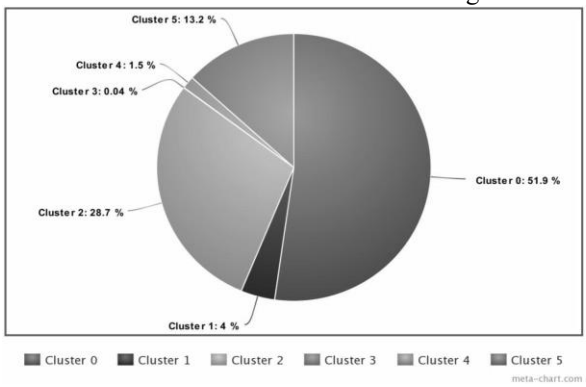

FIGURE 1 – PIE CHART OF CLUSTERS

As seen in figure 1, the most of the instances in Cluster 0 (51.9%).

V. B. Regression Experiments Results

Tables 4, 5, 6, 7 shows a subset of regression experiment results. For each experiment, mean squared error, mean absolute error and R-square score are calculated as mentioned in experiment setup.

LINEAR REGRESSION			
Experiment	**Mean Squared Error**	**Mean Absolute Error**	**R-Square Score**
train_test_split with test_size = 0.4	111147.67	202.80	0.69
train_test_split with test_size = 0.5	110444.71	202.64	0.69
k_fold with k = 5	98067.48	200.65	0.72
k_fold with k = 6	94775.70	200.73	0.74
k_fold with k = 7	94797.43	199.80	0.73
k_fold with k = 8	91375.94	200.32	0.74
k_fold with k = 9	91951.43	201.23	0.75
k_fold with k = 10	98892.48	201.82	0.75

** With K-Fold method, best resulting model selected accross k splits on data.*

Table 4 – Linear Regression Results

POLYNOMIAL REGRESSION				
	Experiment	**Mean Squared Error**	**Mean Absolute Error**	**R-Square Score**
	k_fold with k = 6	59671.49	146.08	0.83
	k_fold with k = 7	60943.03	149.85	0.83
degree = 4	k_fold with k = 8	61908.80	148.05	0.82
	k_fold with k = 9	55186.84	145.55	0.83
	k_fold with k = 10	56228.49	144.93	0.83
	k_fold with k = 6	59656.65	148.42	0.82
	k_fold with k = 7	61693.03	145.82	0.82
degree = 5	k_fold with k = 8	58788.70	146.83	0.84
	k_fold with k = 9	60150.97	146.99	0.83
	k_fold with k = 10	56071.21	147.59	0.82

** With K-Fold method, best resulting model selected accross k splits on data.*

Table 5 – Polynomial Regression Results

STOCHASTIC GRADIENT DESCENT			
Experiment	**Mean Squared Error**	**Mean Absolute Error**	**R-Square Score**
train_test_split with test_size=0.3 alpha=1e-4 max_iter=1000	2,78E+46	1088728319845761.9	-7,63E+38
train_test_split with test_size=0.3 alpha=1e-3 max_iter=1000	1,35E+46	915178551426498.8	-3,70E+40
train_test_split with test_size=0.3 alpha=1e-2 max_iter=1000	5,67E+47	1,56E+32	-1,56E+42
train_test_split with test_size=0.3 alpha=1e-1 max_iter=1000	3,52E+46	1396274709447520.8	-9,66E+39
train_test_split with test_size=0.3 alpha=1 max_iter=1000	2,99E+46	1,14045E+15	-8,22E+39

Table 6 – Stochastic Gradient Descent Results

LINEAR SVR			
Experiment	**Mean Squared Error**	**Mean Absolute Error**	**R-Square Score**
train_test_split with test_size=0.3 C=1e-2 max_iter=1000	105378.68	197.39	0.69
train_test_split with test_size=0.3 C=1e-1 max_iter=1000	192546.62	343.68	0.44
train_test_split with test_size=0.3 C=1 max_iter=1000	121095.43	201.49	0.65
train_test_split with test_size=0.3 C=10 max_iter=1000	121095.43	201.49	0.65
train_test_split with test_size=0.3 C=100 max_iter=1000	121095.43	201.49	0.65

Table 7 – Linear Regression Results

Linear regression and polynomial regression models worked well, however polynomial regression results better, naturally. Stochastic gradient descent gave unreasonable results most probably algorithm requires more number of iterations, it couldn't converge to the optimum solution for data of this study. Support vector regressor results are not so bad, but worse than polynomial and linear regressions'. Support vector regressor couldn't converge too, it requires

more number of iterations. Since the data contain around 85.000 instances and the support vector regressor's fit time complexity is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10.000 samples[4], its normal to fail to converge of support vector regressor (SVR).

According to derived equation (1), best model is chosen as a polynomial regression model with a polynomial degree of 5.

CONCLUSIONS

Features	**Prediction - Total Sales Amount on Given Day**
Product 50 - 15/04/2019 Inflation change: +4,0%	733
Product 50 - 15/04/2019 Inflation change: -4,0%	1160
Product 99 - 04/07/2019 Inflation change: +9,0%	-403
Product 77 - 21/12/2019 Inflation change: -4,0%	1981
Product 77 - 21/12/2019 Inflation change: +4,0%	1629
Product 56 - 23/09/2020 Inflation change: +5,6%	636
Product 56 - 23/09/2020 Inflation change: -5,6%	3057

TABLE 8 – PREDICTION RESULTS

The table 8 shows the predictions for certain products using polynomial regression method with polynomial degree of 5 which is chosen according to regression experiments.

Different interpretations can be made using the sales difference, seasonality and inflation. Product 50 and product 77 does not react too much with the inflation change, we can conclude that these products may be a must for a household so the consumer continues to buy these products even with the inflation change. Prediction on product 99 with positive inflation change results with a negative sales figure. This basically means that the company should not sell this item when there is that much of a change in inflation. Finally even with high inflation rate product 56 is not on negative side. This means that even with high inflation rates some products still manage to survive but it's nearly impossible for any product not to react according to the inflation rate.

REFERENCES

[1] https://www.effectmanager.com/forecast-budget
[2] https://www.capterra.com/p/157977/SalesLoft/
[3] https://help.zoho.com/portal/kb/articles/get-started-introduction-zoho-crm
[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html