Forecasting sales based on card transactions data

Alexandra Moraru, Dunja Mladenić
Jozef Stefan Institute and Jožef Stefan International Postgraduate School,
Jamova 39, 1000 Ljubljana , Slovenia
firstname.lastname@ijs.si

ABSTRACT

Smart cities are an important topic in today's research problems, with high impact in many domains from economy to transportation, health and living style. The problem addressed in this paper is that of sales forecasting for a specific category of products. We present the results of three regression algorithms, applies on real live data, for predicting the cumulative hourly sales of petrol. The prediction is made for three short term intervals, of 1, 4, and 8 hours into the future. A study has also been conducted in order to identify the amount of historical data required for optimal results.

Keywords

Data mining, sale forecasting, regression algorithms, smart cities.

1. INTRODUCTION

Smart cities are an important topic in today's research problems and can be defined as complex problems, combining multimodal data from several sources. The high level requirements for making a city smarter, as envision by IBM in the larger Smarter planet program [1], refer to the collaboration and coordination between city agencies managing different domains (e.g. water management, transportation, buildings, etc.) in order to be able to optimize the limited resources and to efficiently and effectively deliver city services. Moreover, different technologies may enable smarter cities, such as: communication channels (e-mail, instant messaging, etc.), business rules, data sharing (data models, accessibility) and integration of different sources of data [4]. In another study [3] the classification of cities as smart is made based on 6 criteria: economy, people, governance, mobility, environment and living. The problem presented in this paper relates to the first criterion, the smart economy, with indirect relation to transportation and living, as the main objective is that of petrol sales forecasting. [2]

The task of forecasting employs a set of methods and tools for making predictions of the future, relying on past and present data and analysis of trends. A well-known example is the prediction of a target variable at a specific time in the future. Sales forecasting is important in business management and decision making. Short-term and long-term sales forecasting can be affected by many factors, including economic up or downturns, changing trends and fashion, season, etc.

A typical research problem is sales forecast for a particular business or industry. In this paper we present the results of cumulative sales forecasting, where the amount we are predicting refers to a whole category of products, over a larger geographical area, independent of the particular businesses performance. More specifically, we predict the petrol sales in a city, given historical card transactions. Individual card transactions are aggregated on an hourly basis and used for short term prediction of 3 different intervals: 1, 4 and 8 hours into the future. Three regression algorithms are applied on a real live datatset, and their performance is evaluated for different forecasting.

The rest of the paper is structured as follows. Section 2 describes the data used in our experiments. Section 3 describes the methods and algorithm used, while Section 4 presents the results. Finally we conclude the paper.

SiKDD'15, October 1–2, 2015, Ljubljana, Slovenia.

2. DATA DESCRIPTION AND PREPROCESSING

The data used in our experimentation consists of individual card transactions over a period of 13 days. The information available is the time of the transaction, the amount of euros spent and the category of goods purchased. We focus on overall petrol purchases in a city and describe this data in detail.

The aggregated amount (sum) of euros spent on petrol every hour is illustrated in Figure 1. Based on this data, our main objective is to predict the hourly consumption for different time intervals in the future, respectively 1, 4, and 8 hours. A second objective is to analyze how much historical data is optimal for our prediction.

A training instance consists of the current hour of transaction and several aggregated amount of euros spent every hour in the past, for various time intervals. For example, if we would like to predict the amount of euros spent in the next hour, using 4 hours of historical data, our training instance consists of 6 attributes: current hour, euros spent this hour, euros spent 1, 2, 3 and 4 hours ago, where the class (target attribute) is euros spent this hour.

The dataset created for experimentation consists of 303 instances, which contain up to 26 attributes (one attribute is the prediction hour, the rest are hourly aggregates), for the experiments where 24 hours of history are used. All attributes are numeric and there are no missing values. The statistic values of 1 hour aggregates euros spent on petrol, for minimum, maximum, mean and standard deviation are presented in Table 1.

Table 1. Statistic measures of 1 hour aggregate consumption for petrol

Statistic	Value
Minimum	0
Maximum	430982.18
Mean	81529.56
Standard Deviation	83710.81

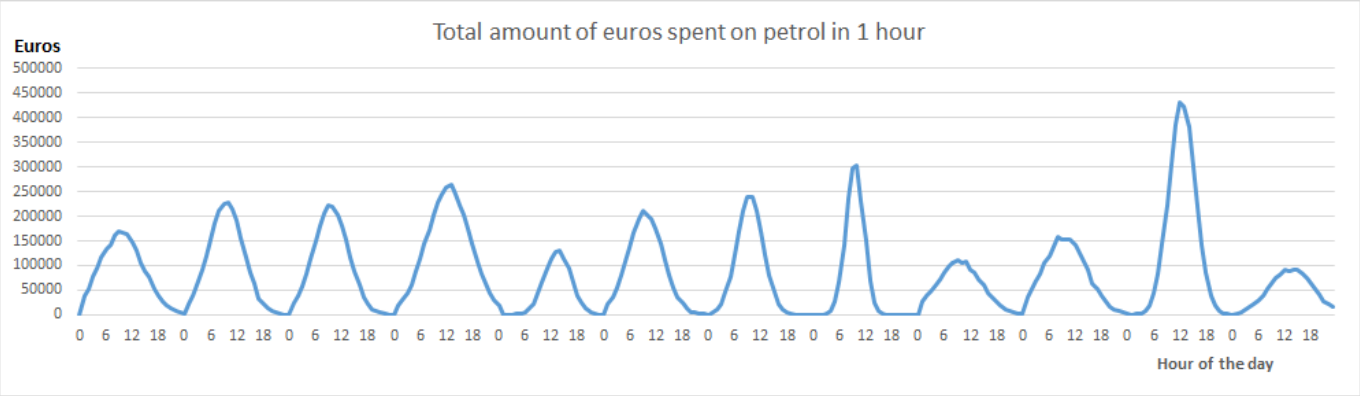

Figure 1 Hourly amount of euros spent on petrol

3. LEARNING METHODS

As our objective is to predict the amount of euros spent in 1 hour for the selected purchase category (petrol), the problem can be formulated as a regression problem. As several regression algorithms are available, we took into account the simplicity, the model understandability and the best reported performances reported, before selecting the algorithms for experimentation. Therefore, for our experiments we have selected three regression algorithms: linear regression, regression tree (M5P model tree and rules) and support vector machine for regression (SMOreg). All algorithm implementation have been selected from Weka toolkit [5].

The target variable to be predicted is the amount of euros that will be sent in 1 hour, 4 hours and 8 hours into the future. We conducted several experiments, providing between 4 and 24 hours of historical data (in 4 hour increments). The performance results of the algorithms are reported in Table 2.

For evaluation we have used separate training and test set, as we consider it to be closer to real live situation, compared to cross-fold validation. The split percentage is 66%, and the order of split is preserved, meaning that first 200 instances are used for training and remaining 103 instances are used for test.

Table 2 Evaluation for prediction of total amount of euros spent in 1 hour for transactions in the petrol category. The two evaluation measures reported are correlation coefficient and relative absolute error. The predictions are made for 1 hour, 4 hours and 8 hours into the future, while the number of hours in the past, used for training the models, are varied between 4 and 24 hours. The algorithms evaluated are support vector machine for regression (SMOreg), regression tree (M5P) and linear regression (LinearReg).

	predicting 1 hour ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.9938	0.9942	0.994	0.9926	0.99	0.9853	10.4874	10.3596	10.9284	12.3995	13.5481	16.786
M5P	0.9623	0.9785	0.983	0.973	0.9652	0.9378	22.5781	20.2358	19.1709	22.8912	23.6852	29.8357
LinearReg	0.5223	0.706	0.5565	0.5237	0.5237	0.8437	80.1759	71.7612	83.5529	85.5285	85.5285	57.3464
	predicting 4 hours ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.7753	0.8165	0.8011	0.748	0.7168	0.6811	58.5435	56.5327	60.8278	67.2914	63.6155	76.0387
M5P	0.7696	0.709	0.6801	0.7346	0.6276	0.6107	60.492	63.7024	73.8301	68.9027	76.6383	88.384
LinearReg	0.7362	0.6628	0.6212	0.5785	0.6327	0.6107	66.589	77.3408	83.7837	90.6461	74.2057	88.384
	predicting 8 hours ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.4957	0.5594	0.5294	0.4599	0.4485	0.4232	100.2677	94.3854	96.3161	96.5257	97.1401	108.3501
M5P	0.5637	0.5939	0.6238	0.4883	0.4919	0.4892	82.5027	88.767	88.5318	103.6142	98.4414	99.3843
LinearReg	0.277	0.5528	0.5131	0.497	0.4527	0.4619	104.3004	98.7528	103.1733	92.3864	101.7982	103.167

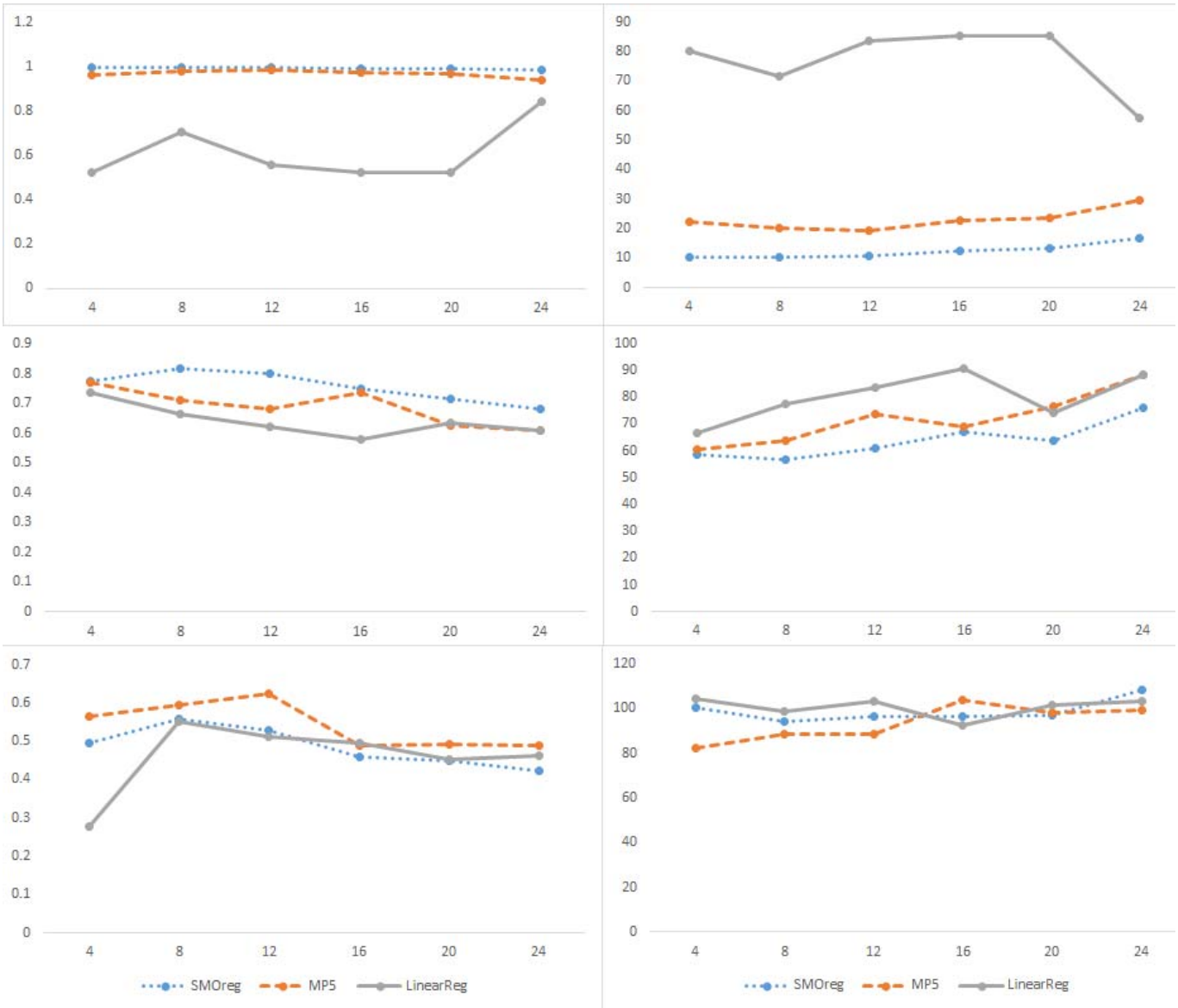

Figure 2 Algorithms performance for different amount of historical data used, reported by the correlation coefficient (left side) and relative absolute error (right side). The top graphs are for 1 hour ahead prediction, middle graphs for 4 hours ahead prediction and bottom graph for 8 hours ahead prediction. On the Y axis are the values of the evaluation measured and on the X axis is the amount of historical data used for building the models (from 4 to 24 hours, in 4 hours increments)

4. RESULTS AND DISCUSSION

The algorithms have been tested for different amount of historical data used in building the prediction model, in order to find the optimal amount of historical data needed and to identify the more robust algorithms for our problem.

The performance of the learning algorithms has been measured in terms of correlation coefficient and relative absolute error.

From the analysis illustrated in Figure 2, several observations can be made:

- As expected, the algorithms performed best when the amount of euros spend predicted is for a shorter time interval in the future.
- SMOreg algorithms performed the best in most cases, closely followed by M5P.
- Shortest prediction interval, that of 1 hour ahead, does not benefit from more the 12 hours of historical data
- Linear regression presented high variability to the amount of historical data provided, counter intuitive to what one would expect. More specifically, it can be observed the when given more the 8 hours of historical data the performance of the algorithm decreases.
- When the prediction interval is larger (8 hours into the future) none of the algorithm reported very good performance, however, it can be noticed that M5P is slightly superior to the rest.

The actual and predicted values for 1 hour petrol consumption are illustrated in Figure 3. The best performing model has been selected for each of the 3 category of prediction: 1, 4, and 8 hours. The first two graphs illustrate the results using SMOreg algorithm for 1 and 4 hours ahead prediction and the third graph illustrated the results for 8 hours ahead prediction using M5P.

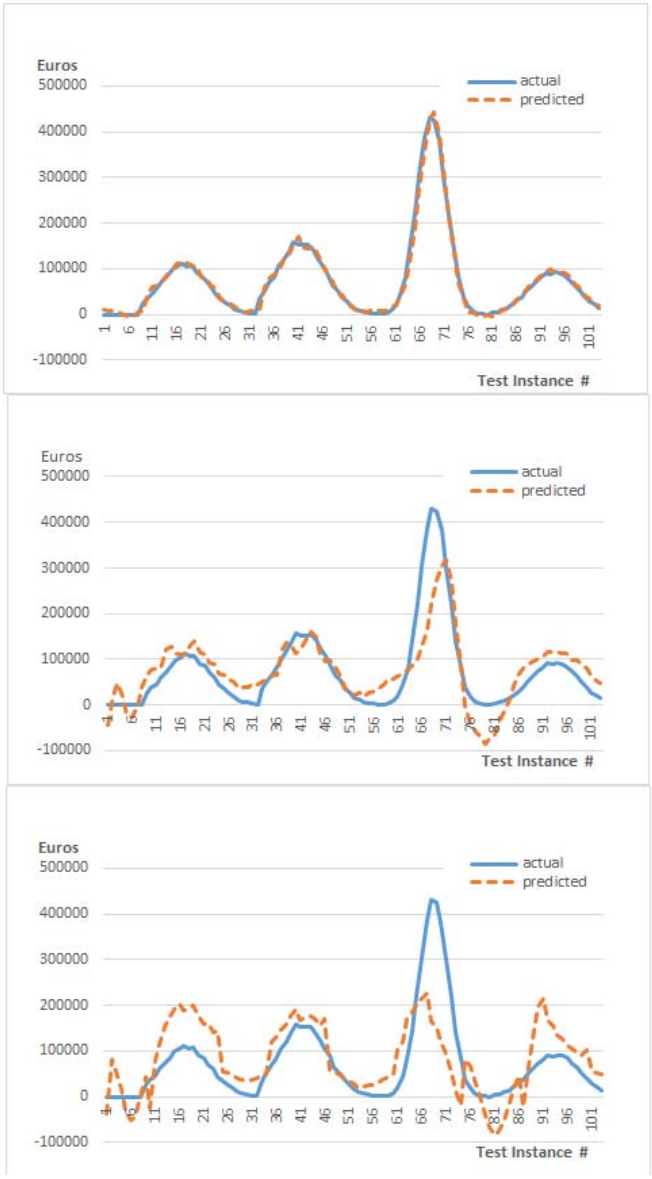

Figure 3 Actual and predicted values of 1 hour petrol consumption. From top to bottom: 1 hour in the future using SMOreg, 4 hour in the future using SMOreg, 8 hours in the future using M5P

5. CONCLUSIONS

In this paper we have reported the results of our study of predicting the amount of euros spent in one hour for a specific purchasing category, using different amounts of historical data. The machine learning algorithm used in the experiments are linear regression, regression tree and support vector machine for regression. The performance measures reported are correlation coefficient and relative absolute error. The results for 1 hour ahead prediction have been very good, while 4 and 8 hours ahead prediction only satisfactory.

Possible improvements for last two cases could be obtained if more than 13 days of data is available. Future work can be conducted in the direction of linking this dataset to other data for the time and geographical region, such as popular event, holidays, transportation.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under PlanetData (ICT-NoE-257641) and NRG4Cast (FP7-ICT-600074).

7. REFERENCES

[1] A Smarter Planet: *http://www.ibm.com/smarterplanet*.

[2] Berry, M.J. a. and Linoff, G.S. 2004. *Data mining techniques: for marketing, sales, and customer relationship management*.

[3] Giffinger, R. et al. 2007. *Smart cities Ranking of European medium-sized cities*.

[4] Wang, Q. et al. 2010. Smarter City: The Event Driven Realization of City-Wide Collaboration. *2010 International Conference on Management of e-Commerce and e-Government*. (Oct. 2010), 195–199.

[5] Witten, I.H. et al. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.