Forecasting sales based on card transactions data

Alexandra Moraru, Dunja Mladenić

Outline

- Introduction
- Data
- Objectives
- Learning Methods
- Results
- Conclusions

Introduction

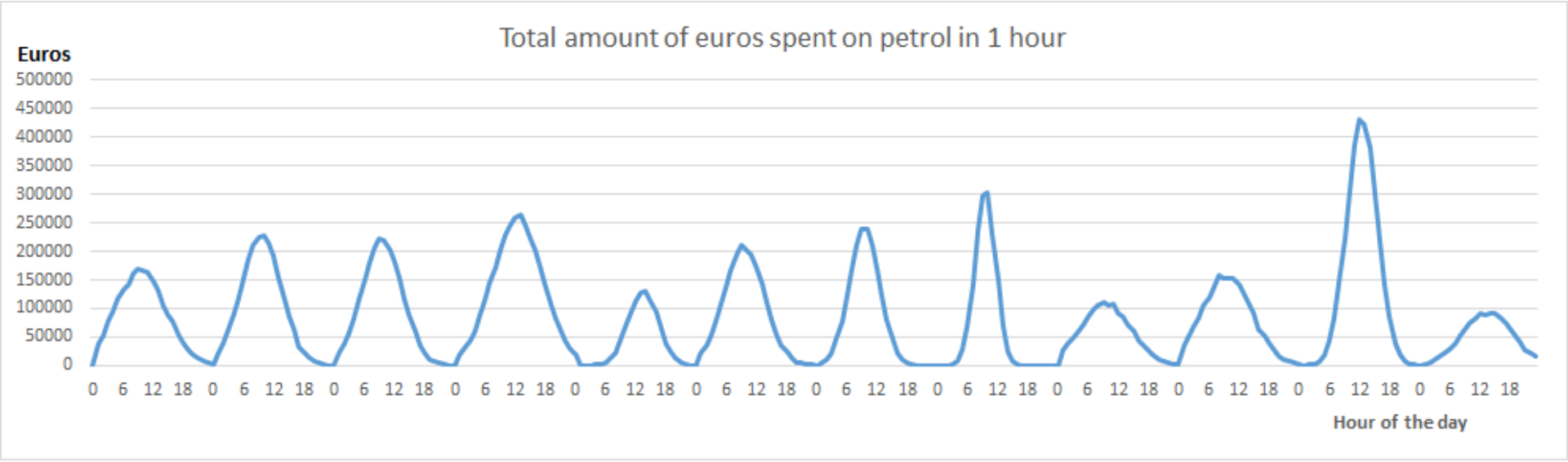
- The task of forecasting employs a set of methods and tools for making predictions of the future, relying on past and present data and analysis of trends.
  - A well-known example is the prediction of a target variable at a specific time in the future.
  - Sales forecasting is important in business 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.
- Cumulative sales forecasting
  - the amount predicted refers to a whole category of products, over a larger geographical area

Data

- individual card transactions over a period of 13 days
  - time of the transaction, the amount of euros spent and the category of goods purchased.
  - focus is on overall petrol purchases in a city

Data

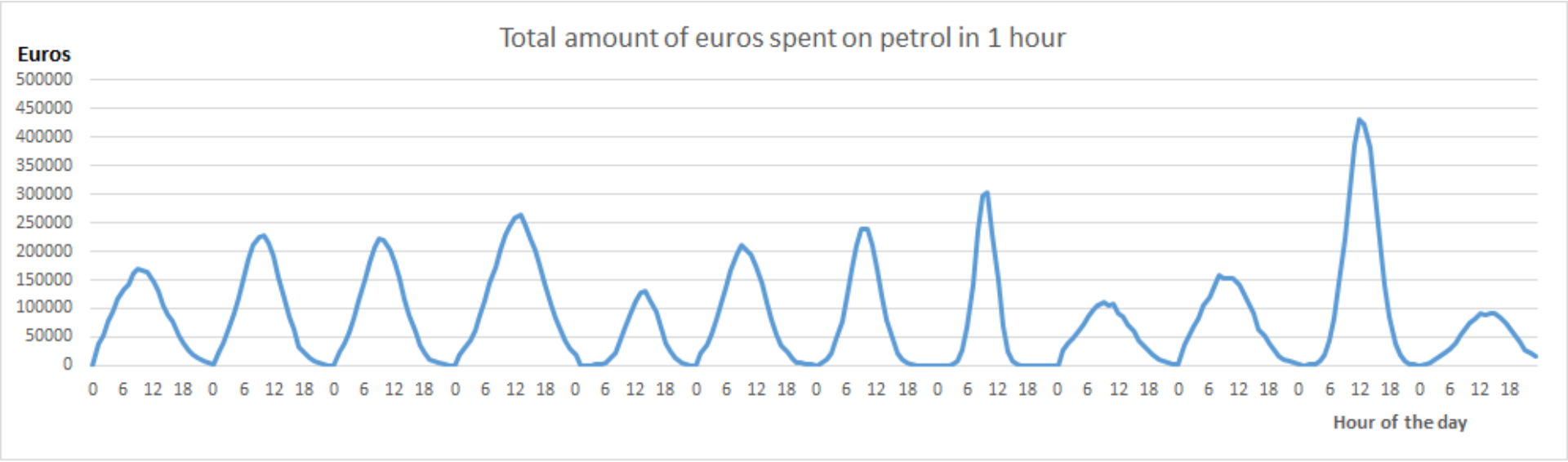
- individual card transactions over a period of 13 days
  - time of the transaction, the amount of euros spent and the category of goods purchased.
  - focus is on overall petrol purchases in a city



Data

- individual card transactions over a period of 13 days
  - time of the transaction, the amount of euros spent and the category of goods purchased.
  - focus is on overall petrol purchases in a city

Statistic	Value
Minimum	0
Maximum	430982.18
Mean	81529.56
Standard Deviation	83710.81



Objectives

- The main objective is to predict the hourly consumption for different time intervals
  - 1, 4, and 8 hours in the future
- A second objective is to analyze how much historical data is optimal for our prediction
  - Experiments run using between 4 and 24 hours of history data

Data Preprocessing

- training instance consists of the current hour of transaction and several hourly aggregates
  - amount of euros spent every hour in the past, for various time intervals.
  - the class (target attribute) is euros spent this hour.
  - 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,
- The dataset created for experimentation consists of
  - 303 instances
  - Each instance contains up to 26 attributes (one attribute is the prediction hour, the rest are hourly aggregates),
  - all attributes are numeric and there are no missing values.

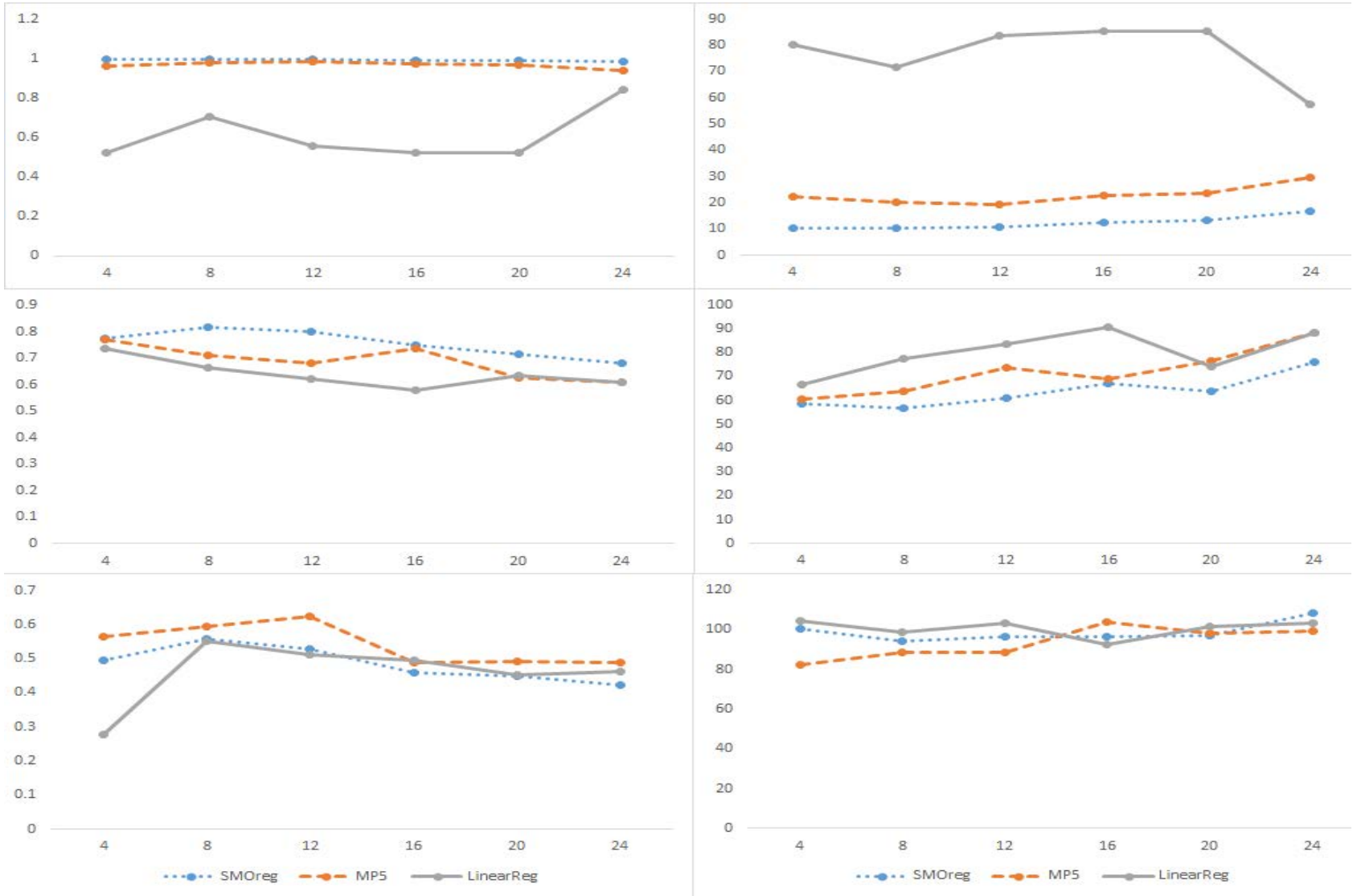
Learning Methods

- Regression algorithms were chosen taking into account simplicity, the model understandability and the best reported performances reported
  - linear regression
  - regression tree (M5P model tree and rules)
  - support vector machine for regression (SMOreg).
  - all algorithms available in Weka
- 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.

Correlation coefficient and relative absolute error

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

Correlation coefficient and relative absolute error



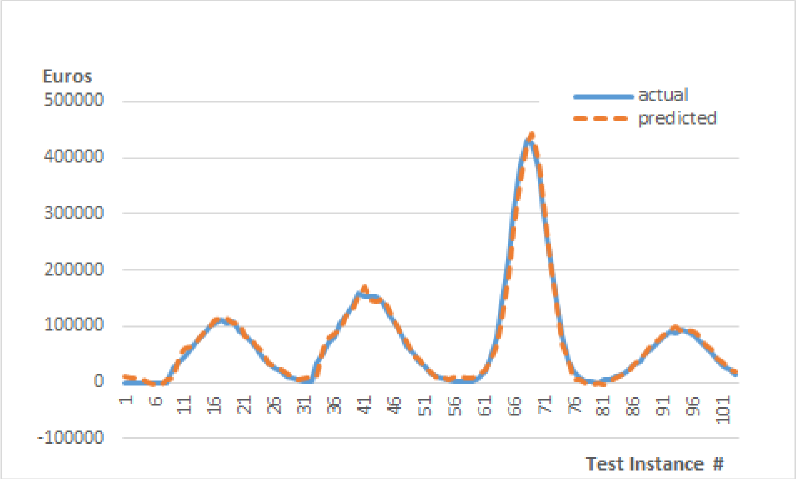
Results discussion

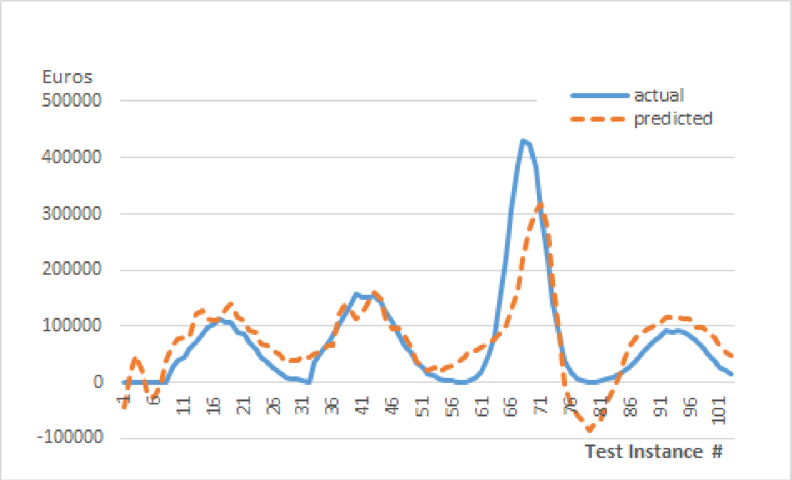
- The algorithms performed best for shorter time intervals
- SMOreg algorithms performed the best in most cases
- Shortest prediction interval, that of 1 hour ahead, does not benefit from more the 12 hours of historical data
- Linear regression presented high variability to the amount of historical data provided
  - 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
  - M5P is slightly superior to the rest.

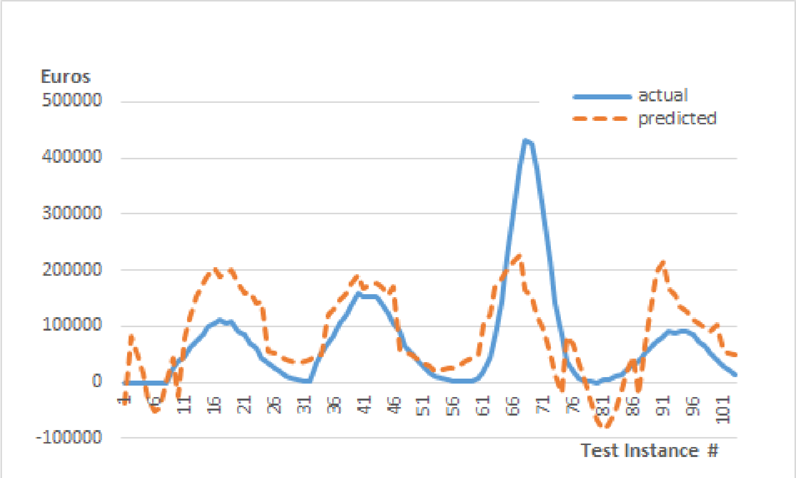
Actual vs predicted values

The actual and predicted values for 1 hour petrol consumption

The best performing model has been selected for each of the 3 category of prediction: 1, 4, and 8 hours.







Conclusions and future work

- We have presented the results of our study on predicting cumulative sales
  - the amount of euros spent in one hour for a specific purchasing category, using different amounts of historical data.
- We analyzed the results of three ML algorithms:
  - linear regression, regression tree and support vector machine
- 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 same time and geographical region
  - popular event, holidays, transportation.