TARTU UNIVERSITY
FACULTY OF SOCIAL SCIENCES

NARVA COLLEGE
STUDY PROGRAM “INFORMATION TECHNOLOGY SYSTEMS DEVELOPMENT“

Kirill Varnatšov

PRICE PREDICTION USING REGRESSION ANALYSIS IN MACHINE LEARNING - A CASE STUDY OF ESTONIAN CHAIN SUPERMARKETS

Bachelor’s thesis

Supervisor: Chen-Wan Lin, PhD

NARVA 2023

PREFACE

For the most of history people are trying to minimise effort and maximise the results of their work, thus knowing the correlation between different events/objects contributes to knowing more about causal relationships around us. Although we already understand correlation in a price changing scope, it is useful to consolidate existing knowledge and find new data in the example of Estonian grocery chain markets. The stores that have been chosen are *Rimi*, *Coop*, *Selver*, *Prisma*, and *Maxima*.

The main aim of this study is to collect the Estonian grocery stores price data and develop Machine Learning models using the data to test their potential efficiency for future works. The reason why the topic of price prediction was chosen, is mainly due to highly volatile price changes in last several years in The Republic of Estonia. This study does not research the reasons of price changes; it does not try to understand the background and specific cases of why the prices change in Estonia.

In this research, I examine correlations, that I think are present, and try to find the unusual ones during this research if any persist. Research results bring us bigger insight of Estonian grocery market price correlations and its overall statistics. Moreover, by the end of the research, the price data for roughly 10 months is collected and stored online.

Finally, my future bigger goal is to develop a working codebase for Linear Regression model training, fine-tuning, evaluating, and using for some other projects and purposes. This is covered in more detail in Chapter 5.2.

CONTENTS

1 INTRODUCTION

Prediction is and continues to be an interesting area of study. It helps us have more certainty in the future and lets us make more precise and weighty decisions. Overall, the prediction is being made based on preceding observed data, because it is the context that matters.

For predicting prices of a grocery store, the area of Machine learning is used for this research. In Chapter 1.1 the brief information about Machine learning is given, and Chapter 1.2 tells about aims and objectives that were raised for this study.

1.1 Machine learning

Machine learning is a Computer Science discipline, that M. I. Jordan and T. M. Mitchell (2015, p.255) tie with two questions: "How can one construct computer systems that automatically improve through experience?" and "What are the fundamental statistical computational-information-theoretic laws that govern all learning systems, including computers, humans, and organisations?"

While machine learning is a quite wide area, regression being a small portion of it. Regression lies into the part of supervised learning, meaning the dataset is labelled. Zhi-Hua Zhou (2018, p.44) in their work "A brief introduction to weakly supervised learning" defines supervised learning as "…techniques construct predictive models by learning from a large number of training examples, where each training example has a label indicating its ground-truth output."

Zhi-Hua Zhou (2018, p.44) also defines labels in regression to be "a real-value response corresponding to the example".

1.2 Aims and objectives

For numerous reasons we would want to either predict upcoming prices of a specific product (or a group of products) depending on how another product (or products) have changed or will change or to find the correlation between different products. It is always useful to know how and at what degree those product correlate to each other (Draper N. R & Smith H., 1966, p.1).

Some of the real-life problems, that can be solved by finding correlation:

- Optimization of actions depending on the correlation data (e.g., if there is low correlation between medicine dose increasing and positive medical result, one would be less likely to take more dosage as it makes little to no sense).
- Correlation data can be used as an intermediate information resource for other kind of research.
- Any kind of data analysis, such as market analysis.
- Price prediction.

In this research, I am trying to predict one product price based on another product or several products prices.

This is a regression analysis problem. Linear regression analysis is a method of finding correlation between variables, however it may be used to predict those variables. Linear regression is a simple, but exceptionally powerful tool, which may help us understand ongoing processes more and find non-trivial correlations.

In this research the goals are to spot ongoing correlations within Estonian grocery stores, train different Linear Regression models based on those correlations, evaluate, and test those models on real life data.

Summarising research questions and objectives:

1. Find correlations.
2. Based on that correlation conduct regression analysis.
3. Train different Linear Machine Learning models.
4. Evaluate models accuracy on a real-life data.
5. Evaluate models usability and assume use cases within a web application.

The paper consists of Introduction, Background research, Data and method, Results and Evaluation, and Conclusion chapters.

- The Introduction provides with the overview of the problem and regression analysis model.
- Background research gives preliminary knowledge about regression analysis and methods of its computations as well as inflation definition and its fundamental causes.

- Data and methods chapter describes how data was acquired and processed, and what methods were used in this research.
- Result and Evaluation chapter tries to explain regression analysis output correlations as well as evaluate and fine-tune those results from the model. Also, in this chapter some use cases for models are assumed.
- The conclusion summarises results and gives comments on the outputs.

To perform a regression analysis, the dataset is needed, thus it was decided to collect data daily starting on 6th March from 5 different Estonian grocery chain markets: *Coop*, *Maxima*, *Prisma*, *Rimi*, and *Selver*.

The source codebase and datasets are available at GitHub webpage at https://github.com/kirillvarn/grocerycomparator-stat

2 BACKGROUND

Background study is essential for better understanding of the study. The importance of understanding inflation and mathematical background of linear regression is somewhat big. For instance, the essential definition of inflation lets us normalise price data more precisely and get rid of noise in datasets.

2.1 Inflation

According to David L. and Michael P. (1975, p. 741) inflation is a process of continuous rising prices or, in other words, a process of continuous drop of money value. It is, however, precise definition, yet it does not tell us about the causes of inflation itself, Helmut Frisch says (1983, p. 9).

H. Frisch (1983, p. 9 - 10) adds up to the definition:

a. "Only when price increases are irreversible can one speak of inflation without qualification."

b. "... inflation does not concern increases in the prices of individual commodities; it refers to an increase in the general price level …"

c. "One should hesitate to label as inflation increases in the general price level at a rate of less than 1 percent per year … This is necessarily a subjective criterion."

Thus, due to the core definition of inflation, it can be assumed that the data must be time-oriented, meaning, the earlier prices come first.

2.2 Linear regression analysis

Draper and Smith explain regression analysis as follows: "…to examine data and to draw meaningful conclusions about dependency relationships that may exist. … fitting the 'best' straight line to given data to relate two variables X and Y…" (Draper N.R. & Smith H., 1966, p. 4).

One of the most common methods of finding linear regression model parameters is least square estimation method (Draper N. R & Smith H., 1966, p.7 - 9). Simple regression line in first order model is denoted:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where β_0 and β_1 are model parameters, and ϵ is a random error (term error) by which Y may fluctuate (Draper N.R. & Smith H., 1966, p. 8).

Multivariate regression first order model for k number of parameters (independent variables) is denoted:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

2.3 Predictions using linear regression

Linear regression analysis can be used to predict values by their respective predictors (independent values). For such purpose, it is essential to estimate the most accurate model parameters (coefficients) and an error term.

Similarly, one of the most common methods for acquiring linear regression parameters for prediction is OLS. According to S. Weisberg (2005, p.21), OLS method is a minimization of residual sum of squares. Residual is a distance from the actual dependent variable to a fitted regression dependent value (Weisberg S., 2005, p.23).

Usually more than two parameters are used in practice (Draper N.R & Smith H., 1966, p. 140).

Linear regression parameters estimate values using OLS are the values, that minimise residual sum of square equation (Weisberg S., 2005, p. 273):

$$RSS = \sum_{i=0}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Where y_i is a i^{th} value of a variable and $(\beta_0 + \beta_1 x_i)$ is the predicted y.

Estimation of β_0 and β_1 parameters is a calculus problem. To estimate those parameters, it is needed to take partial derivatives of *RSS* equal to 0 with respect to the parameters β_0 and β_1, accordingly:

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = 0$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = 0$$

Which simply means function $RSS(\beta_0, \beta_1)$ slope of a tangent line is equal to 0, what indicates the minimum of the function in this case.

So, parameters for simple linear regression are:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Where $\overline{x}$ and $\overline{y}$ are the mean value of x and y respectively.

2.4 L1 and L2 regularization

Regularization is simply an additional attribute to the Linear Regression model, that penalises the overly complicated model. From Demir-Kavuk et. al. words, L1 regularization "… adds the sum of the absolute values of the model parameters to the objective function …" and L2 regularization "… adds the sum of the squares …" (Demir-Kavuk, O. et al., 2011)

L1 regularization:

$$OLS + \lambda|\beta_j|$$

L2 regularization:

$$OLS + \lambda\beta_j^2$$

Where λ is just a tuning coefficient. When λ is 0, a simple OLS is being used as a cost function, and when λ is getting bigger, then bigger penalty is being applied to a cost function.

Arthur E. Hoerl and Robert W. Kennard (1970, p.55) also mark: "…that parameter estimates based on minimum sum of squares have a high probability of being unsatisfactory, if not incorrect…" and Robert Tibshirani (1996, p.267) adds to it "… the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients."

3 DATA AND METHODS

This analysis uses daily price data from five Estonian chain-markets, namely *Coop*, *Prisma*, *Selver*, *Rimi*, and *Maxima*. Those stores were chosen, because all of them have an online page with all of the product with necessary data.

The data and its acquisition methods are described in chapter 3.1; data analysis is described in chapter 3.2; model evaluation is done in chapter 3.3.

The entire process is available in a Jupyter notebook at GitHub: https://github.com/kirillvarn/grocerycomparator-stat/blob/main/prod.ipynb.

3.1 Price data

This research uses inferential analysis of Estonian chain-market prices over the course of nearly 8 months starting from 6th of March 2022 and finishing on October 11th.

The data was collected by scraping from beforementioned Estonian chain-markets e-shop internet pages using *python* web-parsing library *beautifulsoup4.* Data is being stored in a relational database *PostgreSQL* with *product name*, *product price*, *shop name*, and *discount* columns in a table of a scraping date as a name.

As of October, the 11st, 79 rows were parsed, meaning prices data from 79 days were collected. Usually, at least 2 parsings per week were performed to keep the data consistency.

As a result of each parsing, in average 40,492 records of various products were received for each parsing and stored in the database. Totally, for 79 parsings 3,198,923 records were acquired.

Since initial data did contain noise, it was essential to normalise given data. It was decided that discounted products do not contribute to better understanding of data analysis, thus discounted products were not considered into final dataset. Discounted prices were extrapolated with prices of previous dates – this method increases robustness and clarity of linear regression analysis. The same method was applied for missing price data – this occurred, when a product went out of stock, so it was considered that price did not change.

All scraped grocery data is available at https://kirillvarn.github.io/grocerycomparator-web/#/archive.

3.2 Choosing the data

This study uses Simple and Multiple regression analysis along with Ridge and Lasso regression to find the best Machine Learning model for price prediction among Estonian chain-market products. Different kinds of products were chosen to be analysed and predicted. The criterion by which variables were chosen is following:

1. Independent variables are unprocessed food – milk, flour, beef meat, pears, and apples.

2. Dependent variables are foods that contain those unprocessed food at some degree – pizzas and cakes, meanwhile having a lot of other ingredients as well.

This set of products was chosen, because of the popularity of the chosen products in Estonia. Milk, beef, pears and apples are highly consumable products that also are used to make a huge variety of other food products. Flour is an essential ingredient that is being used almost everywhere.

Pizza is chosen, because it is highly customisable food and is popular choice among people to eat. Cakes were chosen as something mostly sweet, that also can be made from a great amount of different ingredients.

Also, as for dependent data, it was decided that more complex products are being chosen, that also contain more other products, that are not being considered in this study. It was done for the sole purpose of evaluating how the most fundamental and common products correlate with more compound ones and to see if it is possible to build machine learning models considering a half-empty context/dataset.

However, there is no significant difference, which products could be used to train a Machine learning models in the scope of this research, because Regression analysis works equally on the majority sets of variables. I have chosen those products because of my personal liking.

In total 8 different parameter combinations were analysed.

Response (dependent)	Predictor (independent) parameters

Every Rimi milk	Every other market milk
Pizza	Beef
Pizza	Flour
Pizza	Beef, flour
Cake	Milk
Cake	Flour
Cake	Flour, milk
Cake	Flour, apple, pear

Table 1. *Regression parameters*

The data is being split into 2 datasets – a train dataset and a testing dataset. The train dataset has 80% of variables and the testing dataset has 20% of the data. Given that the data is time oriented, training data comes before the testing data with the respect to time.

3.3 Choosing the models

Sarker (2021, p. 2 - 3) points out four machine learning categories, namely *Supervised learning, Unsupervised learning, Semi-Supervised learning,* and *Reinforcement learning.* For this research, the Supervised learning is the best fit, because as per Sarker: "… It uses labeled training data and a collection of training examples to infer a function …" (2021, p. 3). In the case of this project, label being the price of a product.

For Supervised learning Sarker (2021, p. 4) points out two main methods – *Classification* and *Regression.* Where Classification from Sarker words works as: "… maps a function (f) from input variables (X) to output variables (Y) as target, label or categories …" (2021, p. 5). Meaning, Classification tries to output discrete values and put data into specific categories.

Regression analysis is defined by Sarker as: "Regression analysis includes several methods of machine learning that allow to predict a continuous (y) result variable based on the value of one or more (x) predictor variables …". Meaning, Regression tries to output value on a scale of some values.

Regression analysis is a best fit for the price data collected from Estonian grocery store, because prices do not fall into any category and are a continuous value.

There are several Regression ML modeling techniques being *Simple Linear Regression*, *Multiple Linear Regression*, *Polynomial Regression*, *Lasso* and *Ridge Regression,* which Sarker specifies (2021, p. 8).

For this research, it is assumed that data is *linear*, thus Polynomial Regression is not considered, and the problem of *non-linearity* is omitted from being tackled. To handle the problem of *over-fitting* and *multicollinearity* Lasso and Ridge Regression are being used, although the feature amount is not big in this project.

As pointed out before, data is assumed to be linear, thereby Simple and Multi Linear Regression is a way to go for this study.

3.4 Data pre-processing

The main statistical and mathematical tools were *python* libraries *numpy*, *pandas*, *seaborn*, *matplotlib*, and *sklearn*.

One essential function was developed and used to define R squared score and to rate the models' efficiency and accuracy. correlate_by_one function performes Simple/Multivariate Linear Regression and Gaussian Process Regression. As a result, json file with set of response – predictor variables with their respective R^2 scores were accomplished for further analysis.

R^2 is determined by following formula (S. Weisberg, 2005, p.24 - 32):

$$R^2 = 1 - \frac{RSS}{SYY}$$

Where:

$$SYY = \sum_{i=0}^{n} (y_i - \bar{y})^2$$

RSS has already been denoted, but it can be expressed in a more compact way by substituting $(\beta_0 + \beta_1 x_i)$ with $\hat{y}_i$, thus the final formula is:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$$

Where y_i is a correct dependent value; $\hat{y}_i$ is a predicted dependent value; $\bar{y}$ is a mean value.

The first step before building up a dataset to feed the models is clearing the noisiness from initial dataset. Following table shows the first model training:

Table 2. *Rimi milk correlation with other market milk*

Response variable	**Mean R^2**
Piim Alma kile 2,5% 1l	1
Toorpiim pastöriseerimata Nopri 1l	0.482758621
Piim kile Rimi 2,5% 1l	0.379310345
Piim Rimi 2,5% 1l	1
Piim Alma 2,5% 1l	0.413793103
Piim Marge UHT 3,2% 1l	0.448275862
Piim Tere 2,5% 1l	0.172413793

Considering the data from Table 2, it can be assumed, that mean accuracy of Simple Linear Regression Model for milk prediction based on Rimi milk is 55% accurate.

Piim Alma kile 2,5% 1l and *Piim Rimi 2,5% 1l* both have mean R^2 of 1, what indicates that 100% of these two variables from the testing dataset can be explained by predictor variables. This is a problem of multicollinearity. Such value seems unrealistic, so for better overview it is vital to see the scatter plot of given variables. Other products appear to have meaningful R^2 score, but it may be a false positive example due to noisy data.

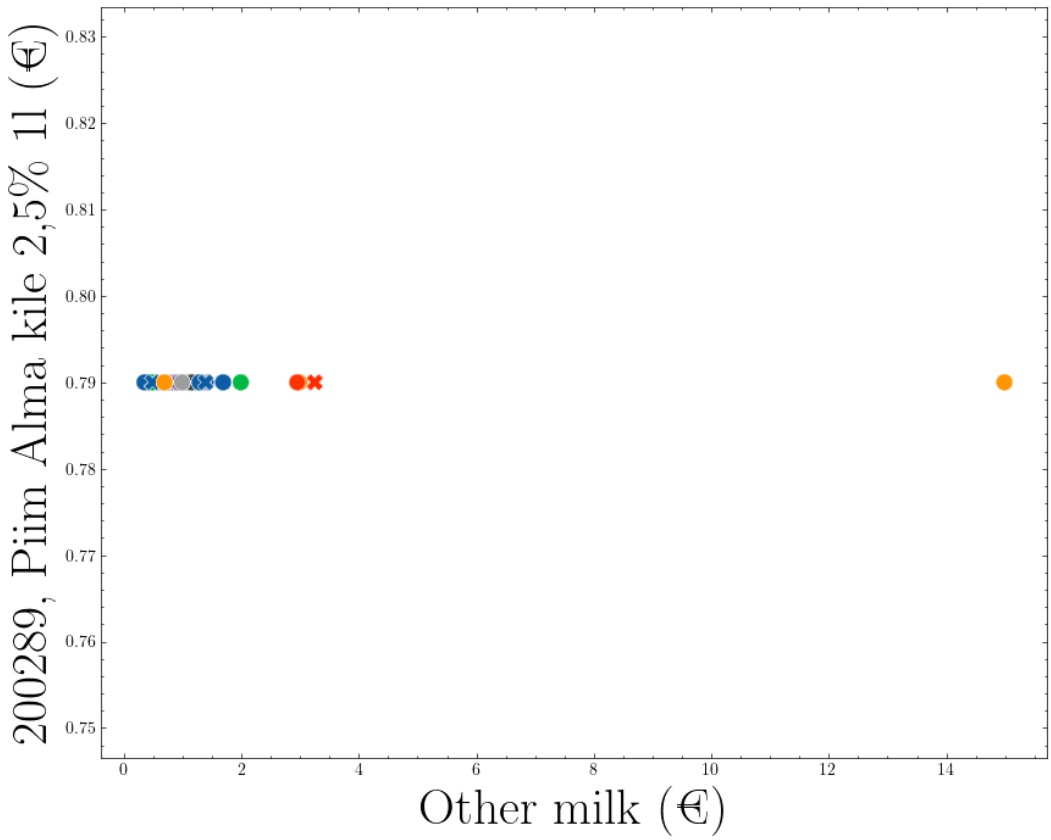

Figure 1. *Piim Alma kile 2,5% 1l* correlation to other market milk

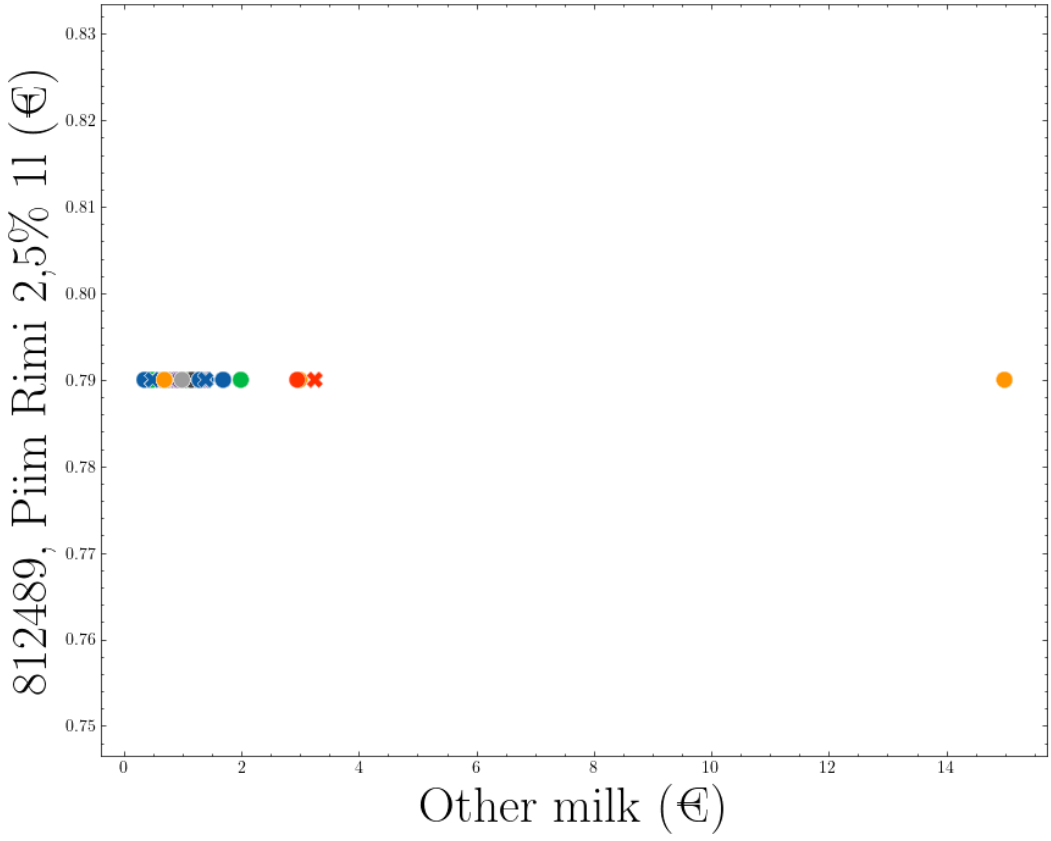

Figure 2. *Piim Rimi 2,5% 1l* correlation to other market milk

It is understood that *Piim Alma kile 2,5% 1l* and *Piim Rimi 2,5% 1l* either have no correlation to other milk or they are underfit to make any viable conclusions. Considering information received from scattered plot, necessary modifications of the trained data were made. Namely, for this research, it is assumed that R^2 value of 1 is simply an underfitting and thus should be omitted.

Table 3 shows the same data as in Table 2 with cleared R^2. N/A means R^2 score was neither 1 nor > 0. Here it is assumed that due to the core definition of inflation, R^2 of negative values are being omitted.

***Table 3**. Rimi milk correlation after data clean-up*

Response variable	Mean R^2
200091, Piim Tere 2,5% 1l	0.0
200104, Piim Alma 2,5% 1l	0.0
263216, Piim Marge UHT 3,2% 1l	0.0
812489, Piim Rimi 2,5% 1l	N/A
804338, Piim kile Rimi 2,5% 1l	0.0
200289, Piim Alma kile 2,5% 1l	N/A
285080, Toorpiim pastöriseerimata Nopri 1l	0.0

Table 3 shows, that the model cannot be used to predict milk prices as its accuracy is 0%. Meaning, milk is excluded from final model building as none of the models show sufficient correlation and thus no Rimi milk can be used to predict other stores milks.

In total 24,011 regression models were acquired using the method of data clearing described above.

3.5 Model building

For model building python's *sklearn* module was used.

Following code piece shows ML regression model building using *sklearn*:

```
x_train, x_test, y_train, y_test = train_test_split( x, y,
test_size=0.2, random_state=0, shuffle=False)
```

```
if model == "lregression":
    ml_model = LinearRegression()
if model == "ridge":
    ml_model = RidgeCV()
if model == "lasso":
    ml_model = LassoCV()
ml_model.fit(x_train, y_train)
```

Firstly, it partitions dataset into train and test data with 80% to 20% relation with no shuffle and random state being zero. It was made due to the fact, that the data is time oriented, meaning we care about what records come first, and what come last.

***Table 4**. Split datasets table amount*

Train dataset	Test dataset
63	16

Test dataset is used to assure how good the model is as well as to get R^2 score values for upcoming predictions, thus this method was used to evaluate regression models.

Secondly, it chooses ML model by the string parameter given to the function, that this code piece is a part of. Lastly, it trains the model. Alpha coefficient is chosen by the *sklearn* cross-validation function automatically based on its performance.

4 RESULTS

This chapter shows regression models comparison results. Here all of the 24,011 resulted models are being tested.

In Chapter 4.1, the resulting models are being split apart by their vectors (dependent – independent variable pairs) and comparing their efficiency. Chapter 4.2 shows models prediction testing and comparing with real life data.

Firstly, it is important to see the overall efficiency of all models. Following table shows 5 topmost models out of 24,011 tested ones:

***Table 5**. 5 topmost regression models by R squared*

Dependent	Independent parameters	R^2

Pitsa salaami ja chorizo Feliciana 320g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise Flank steik 400g	0.95
Pitsa salaami ja chorizo Feliciana 320g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise Flank steik 400g	0.95
Pitsataigen Eesti Pagar 2x300g	Nisujahu 00 Il Molino Chiavazza 1kg; Veisepraad kondita, KARNI, kg	0.95
Õuna-rabarberikook ELT 800g	Piim 2,5%, 1 l, ALMA; Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	0.95
Singi-ja šampinjonipitsa külmut. 425g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise välisfilee steik 360 g, RAKVERE	0.95

The first and the second models were trained using plain Linear Regression; the third, the fourth, and the fifth were trained using Lasso Linear Regression. For more insight about the result, it is vital to see the score distribution. Following figure shows all model distribution of their corresponding R^2 score.

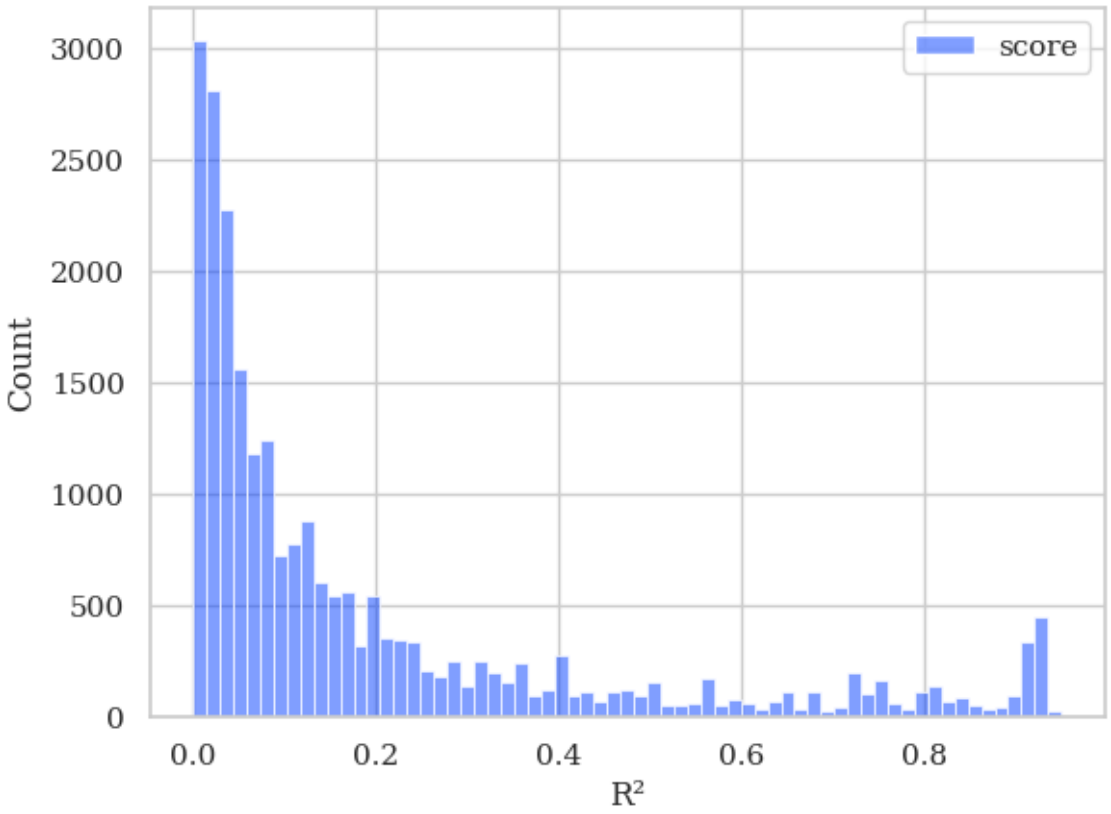

***Figure 4**. Histogram showing occurrences of R squared of all models*

It is apparent, that most of the trained models do not have satisfying level of correlation and therefore cannot be used for prediction purposes. However, there are some of the models showing good level of correlation and thus could be used to predict prices.

For even better overview of how different models perform, it is important to see corresponding models R squares distribution. Following Figure 5 shows this:

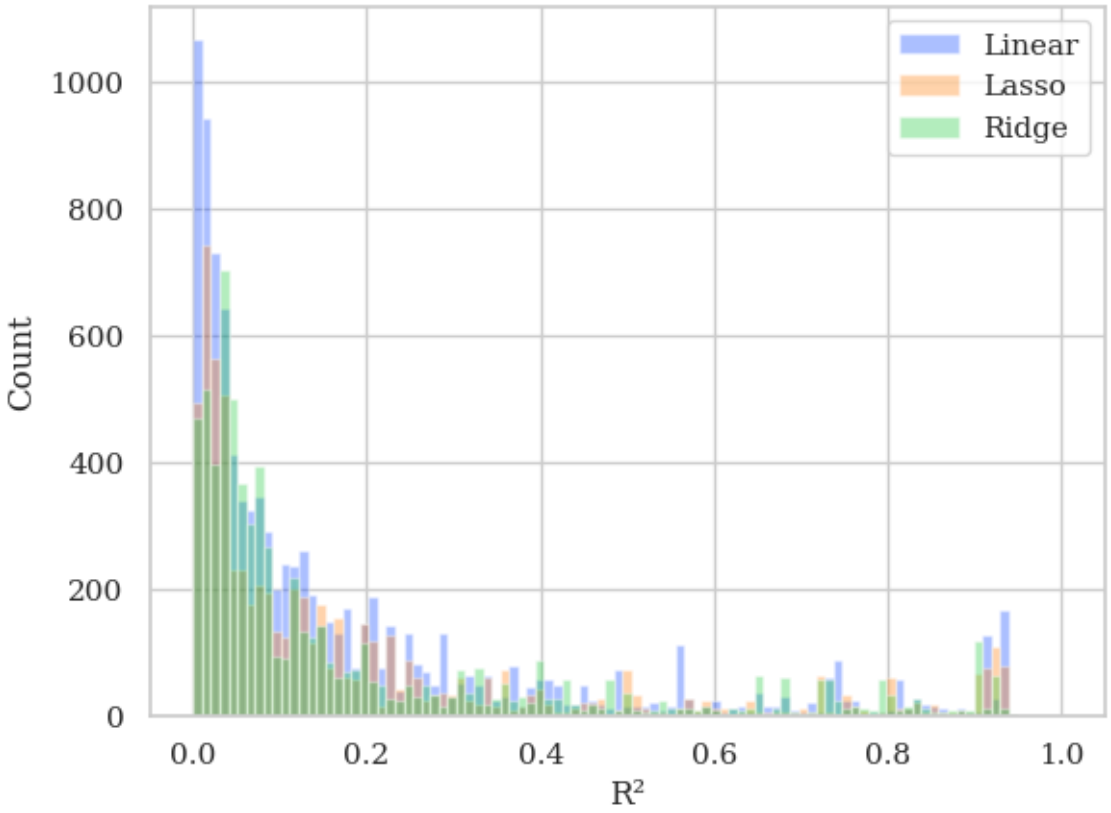

***Figure 5**. Histogram showing distribution of different models.*

All of 3 Regression methods show pretty much the same distribution. However, there are significantly more Linear models than Ridge and Lasso models. In addition, Linear models show more middle scored and high scored models.

There are 7175 valid Ridge models, 6836 valid Lasso models, and 10,000 valid Linear models. Such a preponderance of Linear models can be explained by the constraint introduced earlier. Meaning, more Linear models fall into the R^2 score between 0 and 1. To understand the results fully, mean and median values should be inspected.

***Table 6.** Mean and median value of R^2 of Linear, Ridge, and Lasso models*

	Mean value	**Median value**
Linear	0.190	0.09
Ridge	0.198	0.08
Lasso	0.204	0.1

Considering all the results so far, it can be claimed that on average Linear models perform the worst and Lasso models perform the best, and Ridge models perform somewhat in the

middle. Even though the Linear models showed more middle and high scored models, the low scored models outweigh this.

4.1 Splitting models

In this chapter all models are being split into 3 groups and data is being viewed separately. Out of 24,011 models only 7175 are Ridge Regression models, 6836 are Lasso, and 10,000 are Linear.

The 5 topmost Ridge models are:

Table 7. *5 topmost Ridge Regression models out of 7175.*

Dependent	Independent parameters	Score
Pitsa salaami ja chorizo Feliciana 320g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise Flank steik 400g	0.95
Pitsataigen Eesti Pagar 2x300g	Nisujahu 0 Manitoba Il Molino Chiavazza 1kg; Lihaveise täissuitsuv. Rannarootsi 240g	0.95
Õuna-rabarberikook ELT 800g	Piim TERE 2,5%, D-vitamiiniga 1L kile; Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	0.94
Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Rännukas rebitud veise ja kimchiga 225 g, JOEL OSTRAT	0.94
Pitsataigen Eesti Pagar 2x300g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Rohumaaveise grill-steik küüslauguõli ja chimichurriga, LIIVIMAA LIHAVEIS, kg	0.94

The 5 topmost Lasso models:

Table 8. *5 topmost Lasso Regression models out of 6846.*

Dependent	Independent parameters	Score
Õuna-rabarberikook ELT 800g	Piim 2,5%, 1 l, ALMA; Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	0.95
Singi-ja šampinjonipitsa külmut. 425g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Veise koot	0.94

Pitsataigen Eesti Pagar 2x300g	Nisujahu 00 Il Molino Chiavazza 1kg; Veisepraad kondita, KARNI, kg	0.94
Õuna-rabarberikook ELT 800g	Piim 2,5% pure, ALMA, 1 l; Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	0.94
Pitsataigen Eesti Pagar 2x300g	Nisujahu 00 Il Molino Chiavazza 1kg; Lihaveise snäkk Linnamäe 85g	0.94

Table 9. *5 topmost Linear Regression models.*

Dependent	Independent parameters	Score
Pitsa salaami ja chorizo Feliciana 320g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise Flank steik 400g	0.95
Pitsataigen Eesti Pagar 2x300g	Nisujahu 00 Il Molino Chiavazza 1kg; Veisepraad kondita, KARNI, kg	0.95
Singi-ja šampinjonipitsa külmut. 425g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Lihaveise välisfilee steik 360 g, RAKVERE	0.95
Singi-ja šampinjonipitsa külmut. 425g	Nisujahu 00 Pitsa Il Molino Chiavazza 1kg; Veise koot	0.95
Õuna-rabarberikook ELT 800g	Piim 2,5%, 1 l, ALMA; Nisujahu 00 Pitsa Il Molino Chiavazza 1kg	0.95

It is apparent from the results that all of the methods share the same products, however different independent variables are present. For instance, *Õuna-rabarberikook ELT 800g* is present in all of the 5 topmost results but one with different parameter being the milk. This can be interpreted as the milk parameter being insignificant or that all of the milk products change equally and at the same time, so it makes zero difference of what milk is used.

4.2 Prediction tests

To conduct a prediction test, 3 topmost models are chosen from each Regression methods results dataset. In total 9 models are being tested.

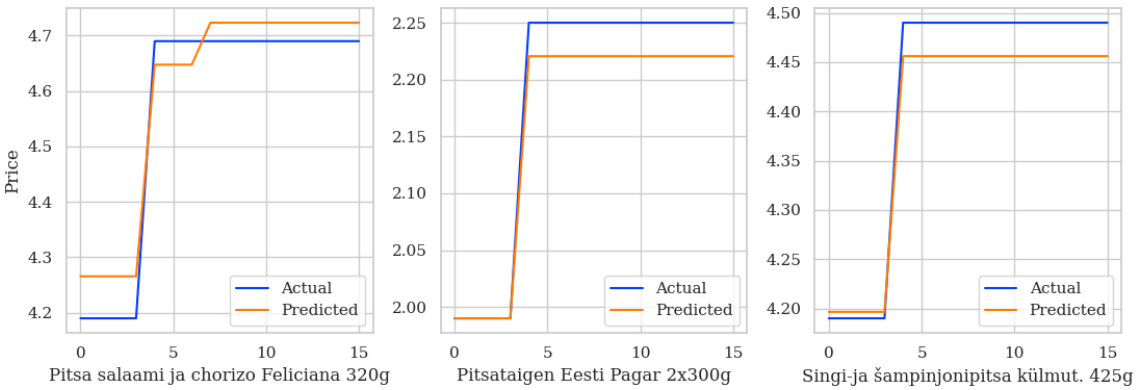

Figure 6. *Linear Regression 3 topmost model comparison with real data.*

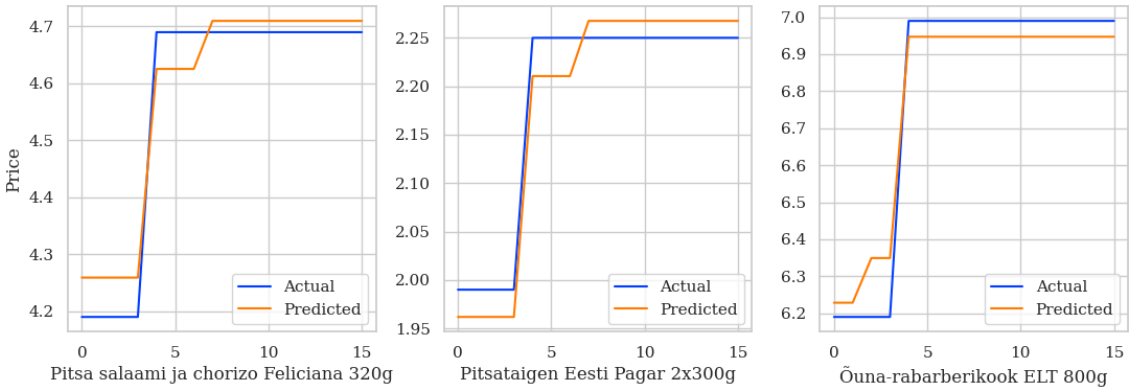

Figure 7. *Ridge Regression 3 topmost model comparison with real data.*

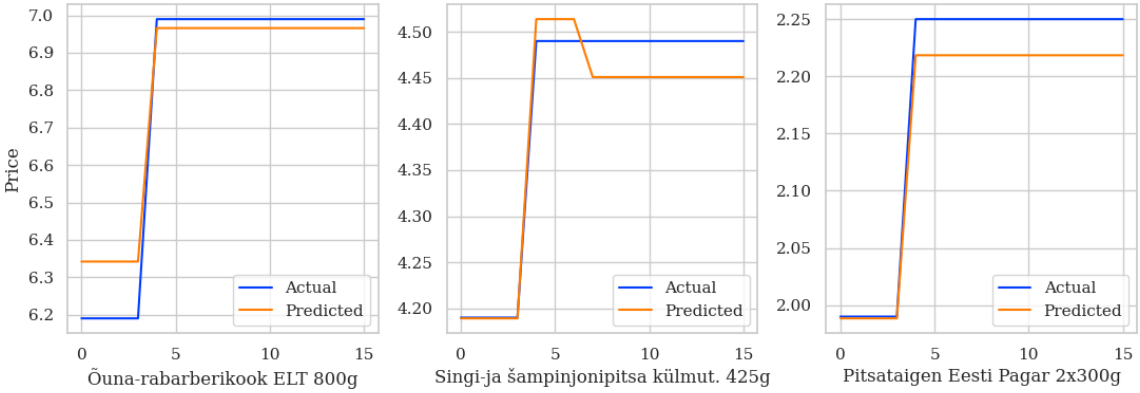

Figure 8. *Lasso Regression 3 topmost model comparison with real data.*

5 CONCLUSTION AND FUTURE WORK

5.1 Conclusion

In the thesis, the problem of choosing and developing best attributes for Regression Machine learning model for further usage in a web application as well as spotting the correlation on the example of Estonian groceries. The results were accomplished by feeding roughly 3 million records to Linear Regression, L1 and L2 regularization Machine learning models. This resulted in 24,011 valid models with $0 < R^2 < 1$.

There are numbers of models that have pretty high R^2, MSE and MAE, which means that current approach can be used to

a) train models,

b) predict Estonian grocery prices based on other groceries using correlation metrics,

if the correlation of the model is high enough.

5.2 Future work

Various ideas could be implemented and tested for the topic of this research. One of my main focuses for the future work is developing an API as well as creating a bigger system in the scope of Estonia grocery prices, containing predictions, statistics, comparing prices, and many others.

As mentioned before, one of my aims in the future is also implementing the code base from this study to predict prices in Estonia based on 1) other products and 2) time.

Some other ideas could be tested:

1. Instead of predicting one product based on another, prices could be predicted in respect to datetime. The web application in question could compute inflation rate based on a basket of products or for a specific product.

2. Other variations of Regression analysis could be performed such as Neural Regression or Decision Tree Regression.

3. Introduce more features for model training.

REFERENCES

Barath. (19.02.2020) Simple Linear Regression for Salary Data. Available at https://www.kaggle.com/code/vivinbarath/simple-linear-regression-for-salarydata/notebook (Last accessed 13.05.2022)

Demir-Kavuk, O. et al. (2011) Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC bioinformatics 12(1), pp. 412–412. DOI: 10.1186/1471-2105-12-412.

Draper, N. R. & Smith, H. (1966) *Applied Regression Analysis*. John Wiley & Sons. DOI: 10.1002/bimj.19690110613

Helmut, F. (1983). Theories of inflation. *Cambridge University Press*. DOI: 10.2307/2232318

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 42(1): 80–86. DOI: 10.2307/1271436

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245): 255–260. DOI: 10.1126/science.aaa8415

Laidler, D. & Parkin, M. (1975). Inflation: A Survey. *The Economic Journal*, 85:741–809. DOI: 10.2307/2230624

Melkumova, L. E. & Shatskikh, S. Y. (2017) Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, 201, pp. 746–755. DOI: 10.1016/j.proeng.2017.09.615.

Sarker, H. Iqbal. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2:160. DOI: 10.1007/s42979-021-00592-x

Seber, G. A. F. & Lee, A. J. (1977/2003) *Linear Regression Analysis, Second Edition*. John Wiley & Sons.

Sudhir Kumar. (04.04.2020) Linear Regression Tutorial. Available at https://www.kaggle.com/code/sudhirnl7/linear-regression-tutorial/notebook (Last accessed 13.05.2022)

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological). 58(1): 267–288.

Weisberg, Sanford. (2005) Applied linear regression. Vol. 528. John Wiley & Sons.

Yu, Chun & Weixin Yao. (2017) Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation* 46.8: 6261-6282. DOI: 10.1080/03610918.2016.1202271

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. National science review. 5(1): 44-53. DOI: 10.1093/nsr/nwx106

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina,

Kirill Varnatšov,
(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
„Price prediction using regression analysis in machine learning - a case study of Estonian chain supermarkets“,
(lõputöö pealkiri)

mille juhendaja on

Chen-Wan Lin,
(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kirill Varnatsov
11.05.2023