Abstract

E-commerce has grown significantly in recent years and more people are choosing to shop online. With the rise of e-commerce, business-related data is also growing, which complicates the traditional pricing of products that have largely been based on manual analyzes. Machine learning and price prediction have been used in various areas to support decision-making. Previous work shows that there is limited research that aims to use machine learning and price prediction of products. The purpose of this research is to bridge this knowledge gap and to contribute to existing research with new data in price prediction of products. The study aims to use a support vector machine during an experiment to predict prices for e-commerce products. The data used contains information about 100,000 purchases made in a real Brazilian online store named Olist, between 2016 and 2018. Furthermore, a grid search is implemented to find optimal hyperparameters. The predictive accuracy is evaluated using two different performance metrics. One of the conclusions drawn from the results of the study shows that the support vector machine can predict prices with good accuracy.

Keywords: machine learning, price prediction, e-commerce, support vector machine, support vector regression

Sammanfattning

E-handeln har de senaste året växt markant och fler väljer att handla på internet. I och med e-handelns uppgång växer också den affärsrelaterade datan vilket försvårar den traditionella prissättningen av produkter som till stor del har grundat sig på manuella analyser. Maskininlärning och prisprediktering har använts inom olika områden för att stödja beslutsfattning. Omfattande litteraturstudier visar dock att studier inom prisprediktering av produkter är begränsad. Syftet med denna studie är att brygga denna kunskapslucka och att bidra befintlig forskning med nya data inom prisprediktering av produkter. Denna studie ämnar till att under ett experiment använda support vector machine för att prediktera priser på produkter inom e-handel. Datan som används innehåller information om 100 000 köp som gjordes mellan 2016 och 2018 i en riktig Brasiliansk online butik som heter Olist. Vidare implementeras en rutnätssökning för att finna optimala hyperparametrar. Den prediktiva tillförlitligheten utvärderas slutningen genom två olika mätsätt. En av slutsatserna som görs utifrån studiens resultat visar att support vector machine kan prediktera priser med god tillförlitlighet.

Nyckelord: machine learning, price prediction, e-commerce, support vector machine, support vector regression

Contents

Terminology
Abbreviations

ML	Machine Learning
AI	Artificial Intelligence
SVM	Support Vector Machine
SVR	Support Vector Regression
RBF	Radial Basis Function
MAE	Mean Absolute Error
RMSE	Root Mean Square Error

Vocabulary in machine learning

For the convenience of the reader frequently occurring words in this study and in the machine learning community are explained.

Table 1: Vocabulary in machine learning and statistics.

Machine Learning	Statistics
Input/Feature	Independent variable
Output/Target	Dependent variable
Training	Fitting
Training set	In-sample set
Test set	Out-of-sample set

Chapter 1

Introduction

Over the last years, e-commerce has become one of the fastest accelerating industries on the market and is today one of the most popular activities taking place on the internet. While many physical stores have suffered from the global outbreak of Covid-19, online retailers seem to have thrived in what has been described as the worst global economy in years. In 2020 alone, e-commerce had 17% of the global retail which was an increase of 3% from the year before (Unctad, 2021). Technological advancements in information technology have supplied customers with the convenience of shopping from anywhere at any given time and day. The availability contributes to better prices which not only benefits the customer but can also increase revenue for a business. According to (Gupta & Pathak, 2014) better prices can be a driving force for increased sales. (Kedia et al., 2020) further suggest that optimal prices in e-commerce is critical to boosting profits.

Product pricing initially relied on human experts manually analyzing data using simple models like linear regression and rule-based methods. The problem with this approach is that it is not scalable. Companies not only had to continuously monitor the process to make sure prices are in line with the market, but they were also costly and sometimes resulted in inconsistent and inaccurate prices. The surge in online shopping has also created an abundance of sales-related data making consistent evaluations challenging. Machine learning could be a useful method that could address the problem of pricing decisions. Machine learning algorithms are capable of processing larger data sets while considering more price impacting factors such as product attributes and historical sales to come up with accurate prices almost instantly. These complex algorithms are also able to find patterns in data which are easily overlooked by humans. This could potentially save decision-makers hours spent on manual pricing and repricing cycles and lead to a more profitable business model (Elraffah, 2021).

The approach of using machine learning to adjust prices of products by predicting product prices is addressed in a study by (Bakir et al., 2018). In their study they experiment with support vector machine and recurrent neural network with long short-term memory to predict prices of mobile phones. The results reveal that support vector machine with a linear kernel is able to predict prices with fewer errors when predicting the price of a single phone model. On the other hand, the authors show that support vector machine predict phone prices with more errors when more phone brands are introduced to the model. In another study, (Johansson, 2018) is implementing different machine learning models to predict prices of vinyl records. The author examines k-nearest neighbors, linear regression, neural network and random forest. Primary findings show that accurate predictions depend on what data is used. The author also discusses how the results are too broad since the whole data set was used and no clear feature selection technique

was presented. Their study concludes that random forest was the optimal model and that better results could have been obtained by putting emphasis on hyperparameter tuning (Claesen & de Moo, 2015). The importance of optimal hyperparameters is further emphasized by (Siwers & Dahlén, 2017) who compares multilayer perceptron, radial basis function network and support vector machine to predict sales in a food store. The authors analyze the variance with ANOVA to determine whether there is significant difference between the models. Their findings show that support vector machine predict food sales with fewer errors than the other assessed models. Differently, (Ihre & Engström, 2019) use k-nearest neighbors and random forest regression. After using grid search to find optimal hyperparameters they discover that random forest predicted house prices with the smallest errors.

Nonparametric machine learning models such as support vector machine and neural network have in the last couple of years gained traction in the scientific community for financial forecasting. According to (Liang et al., 2009) nonparametric machine learning models have an advantage over parametric methods in that they are capable of finding patterns in nonlinear data which results in better prediction accuracy and more efficient decision-making. Support vector machine can also allow for errors within a margin of tolerance which makes it more flexible than simple linear regression models (Bakshi, 2020). The strengths of support vector machine are further demonstrated by (Yu et al., 2013) who is comparing support vector regression and linear regression for newspaper sales forecasting. Their results show that predicted sales with support vector regression are deviating significantly less from the actual sales compared to their linear regression model.

Using machine learning and price prediction is approached by many researchers. Some researchers evaluate different models to predict prices for multiple products whereas some are more concerned with specific products. However, from the reviewed literature it can be concluded that:

- There is limited research that aims to use real e-commerce data and support vector machine to predict product prices.
- The results from the reviewed studies suggest that support vector machine might have potential in the area of e-commerce and price prediction of products.
- Previous studies also shows that support vector machine is able to generalize well on nonlinear data.

Therefore, this study takes a quantitative approach to measure the performance of support vector machine for predicting product prices. The purpose is to bridge the gap of current knowledge and to contribute existing research with new data. This leads to the following research question:

- How well can support vector machine (SVM) predict prices of products using real-world e-commerce data?

The remainder of this study is organized as follows: Chapter 2 will discuss the concept of machine learning and support vector machine. In chapter 3, the methodological process is demonstrated. Then, in chapter 4 results and a discussion is presented. Finally, chapter 5 concludes the study with a conclusion and suggestions for future work.

Chapter 2

Machine learning

Machine learning (ML) is a subfield in artificial intelligence (AI). Machine learning is the study of algorithms able to learn from underlying patterns within data to solve predictive tasks (Abellera & Bulusu, 2017). More specifically, machine learning algorithms can automate decision-making by extracting knowledge from known data that to generalize to novel data (Awad & Khanna, 2015; Müller & Guido 2017).

Machine learning can be further be divided into two learning techniques: supervised learning and unsupervised learning. In supervised learning a predictive model is trained on known input/output values which then can predict the output values on new data (Awad & Khanna, 2015). Supervised learning can be used to solve different problems, where classification and regression problems are arguably the most popular. Classification is used to predict the outcome of discrete output values in form of classes. On the other hand, price prediction is a regression problem and is used to predict the outcome of continuous numerical output values (Sharda et al., 2012). Differently, unsupervised learning aims at finding patterns in data where the output is unknown. Two techniques involve clustering and dimensionality reduction (Vanderplas, 2016).

2.1 Testing and evaluating

Machine learning models are prone to overfit with respect to training data. Overfitting occurs when a model is trying to learn from the underlying relationship between the input and output values along with the noise in training data. As such, the model is too complex and is not able to generalize well to new data. One way of identifying if a model is overfitting is to partition data into a training set and test set. The training set consists of known input/output pairs that is used to fit the model. The test set is a smaller disjunct subset which only holds the output values and is used to evaluate the performance of the trained model (Vanderplas, 2016). A better way of evaluating the performance of a model is to use k-fold cross-validation. It splits the training data into pre-specified number of subsets known as folds. For each fold, one of the subsets is used as validation set while remaining data is used to train the model. The performance is then evaluated on the validation set. Cross-validation is used to give a more reliable estimate of how the model will perform in general because the whole data is utilized as training set and test set (Müller & Guido 2017).

2.2 Support vector machine (SVM)

Support vector machine (SVM) is a nonparametric supervised learning method used for classification problems and regression problems. The essence of SVM is to create a hyperplane which can be described as a line. It is the decision boundary that linearly separates data into disjunct parts. The goal is to maximize the margin between the hyperplane and support vectors on each side of the hyperplane. Support vectors are the data points closest to the hyperplane that support the decision boundary which is illustrated by the triangles in figure 1. In classification the hyperplane attribute data points to different classes depending which side of the hyperplane the data points are (Sharda et al., 2014; Awad & Khanna, 2015).

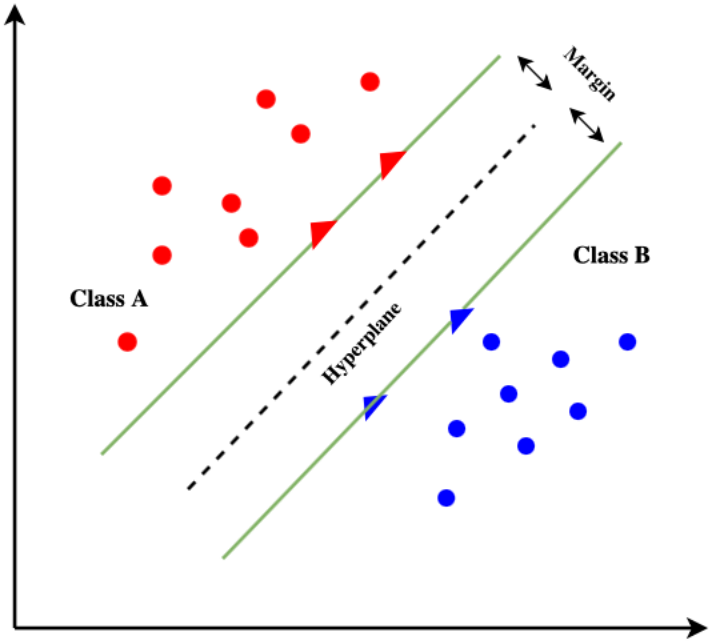

Figure 1: Illustrative example of a classification problem assuming linear separability.

Until now it has been assumed that data is linearly separable. However real-world data is often randomly distributed. To solve nonlinear problems SVM uses a kernel function to transform nonlinear data points into a higher-dimensional space which allows for linear separability (Sharda et al., 2014; Yadav, 2018). Soft-margin SVM further improves nonlinear problems by tolerating error of misclassified data points. It defines a slack variable ξ which measures the extent of violation over linear separability. The degree of violation can be determined by the regularization parameter which will be further discussed in Section 2.3 (Sharda et al., 2014; Awad & Khanna, 2015).

2.3 Support vector regression (SVR)

In this study the objective is not to use SVM for classification but to predict the output value of a continuous numerical value. SVM applies its generalization properties to regression problems but with some minor differences. Instead of looking for maximum separability, SVR defines an epsilon tube that can fit the most data points within its region while minimizing the slack variables. This can be controlled by adjusting the width of the tube. A wider tube is obtained by increasing the value of epsilon allowing more data points to fit within the tube while reducing the slack variables. Decreasing its value will instead narrow the tube resulting in more data points around it. Figure 2 illustrates how the slack variables deviate from the epsilon tube. They are measured based on the distance between the observed data points and the boundary of the epsilon tube. Ultimately, SVR does not consider errors as long as they are within the tube. The tolerance of errors is as already mentioned determined by adjusting the value of the regularization parameter. It penalizes data points that lie outside the epsilon tube. This can be looked at as the tradeoff between the tube width and slack variable minimization. Similar to SVM for classification, one can use a kernel function to map a nonlinear problem into a higher dimensional space. In this study radial basis function (RBF) is used (Smola & Schölkopf, 2004; Awad & Khanna, 2015).

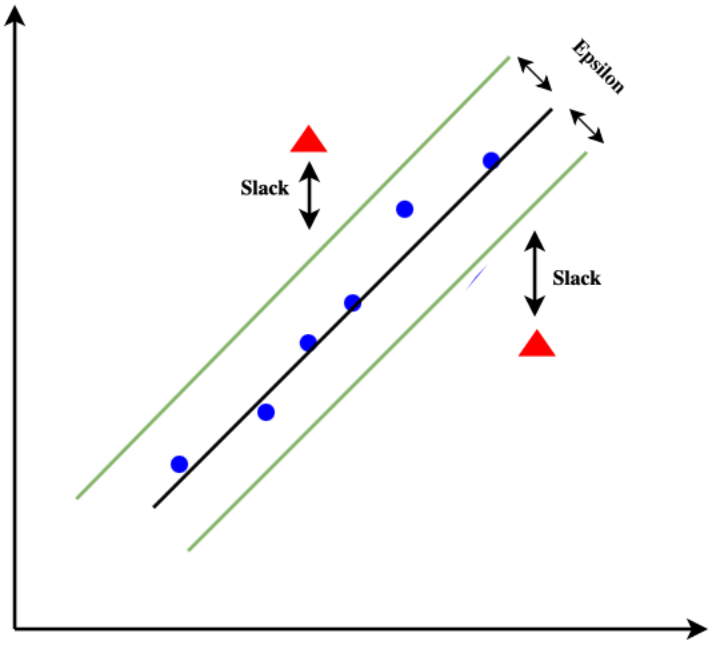

Figure 2: Illustrative example of support vector regression.

2.4 Hyperparameter tuning

Hyperparameters are a set of defined values that are independent from a machine learning model itself and can be tuned to control the process of which a model learns. They are values that control the model complexity and are set before the model is trained. One can think of hyperparameters as the control buttons on a radio that can be tuned to obtain better signal. But finding the optimized hyperparameters by trial and error is time-consuming. In this study a grid search algorithm is implemented with the radial basis kernel function (RBF). Grid search is a hyperparameter technique that defines a grid of unique settings of values that will be tested on the model. It works by exhaustively trying all combinations of values to determine the optimal combination using cross-validation (Géron, 2019).

2.5 Performance metrics

To assess the performance of the proposed model, mean absolute error and root mean square error will be used to compare predictions to actual data. Both metrics have similar criterion where small values indicate that the model is near the real price.

Mean absolute error (MAE) is a metric that computes the mean of the difference between the actual values and the predicted values. However, it does not consider the direction of differences because into only calculates the absolute values. That means that if an error is a negative value, it will only take the positive value into account. MAE can be formulated as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{y}_i - \hat{y}_i| \,,$$

y_i denotes the actual values of the instance *i* and $\hat{y}_i$ the predicted value in that instance. The value *n* corresponds to the number of instances.

Root mean square error (RMSE) will square the errors between actual values and predicted values and then take the root of the average. RMSE can be defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i - \hat{y}_i\right)^2} \,.$$

Both metrics can range from zero to infinite. The difference is that MAE will give equal weight to all the errors whereas in RMSE all errors are squared before the errors are averaged. This means that RMSE penalize larger errors as opposed to MAE which might not be desirable when there are outliers in the data set.

Chapter 3

Methodology

This chapter describes the methodological process of the study.

3.1 Research approach

This study took quantitative approach to answer the formulated research question. A quantitative research strategy involves collecting and analyzing numerical data to find patterns from which conclusions can be drawn. It is also the primary type of data that is derived from experiments. The scale of a research determines the complexity of quantitative technique. It can range from simple quantitative analysis used in smaller projects, to including surveys in large-scaled projects. The typical experiment is characterized by observing or measuring outcomes and to find causality between factors (Oates, 2006).

The chosen research strategy is an interchange between theory and research from surveyed literature. The intent was to gain better understanding of previous work and theoretical concepts. These have been found through open articles, books, and scientific papers. Relevant publications have been accessed through open articles, Science Direct, Springer Link, DiVA, Researchgate and libraries.

3.2 Data collection

Before the experiment began, a complete data set with historical sales transactions had to be collected to fit the predictive model. In this experiment a publicly available real-world e-commerce data set was collected from Kaggle (Kaggle, n.d). It contained information on ~100 000 orders that were made between 2016 and 2018 at a Brazilian online store named Olist. A detailed description of the complete data set can be found in Appendix B. The collected data had initially been stored in different database tables but came in eight different CSV tables during the collection. Each CSV table contained different information about the orders such as customer data and product attributes.

3.3 Data preprocessing

The first preprocessing step was to merge the eight CSV tables. There were in total of 116 581 records and 37 features with a combination of categorical and nominal features. After merging all data into one complete data set, the following step was to perform exploratory data analysis to get valuable insights about the data. It was discovered that many orders had multiple products of which all had to be dropped to avoid any misleading predictions. Many of the features were also consisting of primary keys and foreign keys that originated from the database tables, thus, they were removed from the data set.

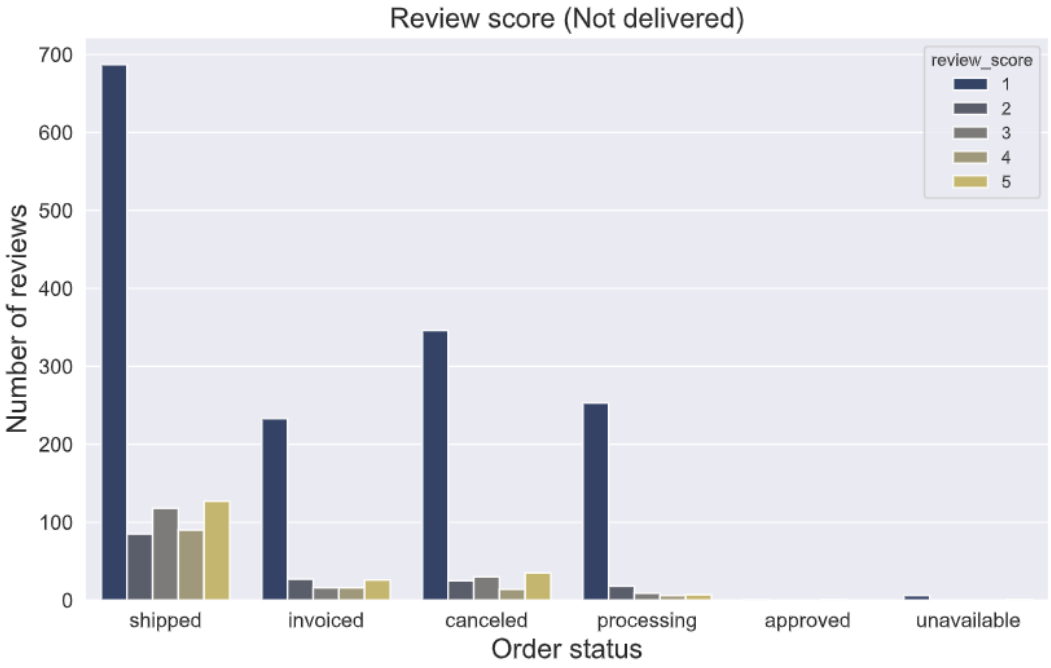

Figure 3: Review score count on order status.

Furthermore, the feature review_score contained ordinal values from one to five indicating the score customers had left on the satisfactory survey regarding their order. The rating scale ranged from one to five. Since it was not specified what the scores meant, it was assumed that one indicated very unsatisfied and five very satisfied. It was also observed that customers were able to review orders before any delivery. This led to a decision to remove any order that had not been delivered to customer. Another preprocessing step consisted of removing prices above 30 Brazilian real partially to reduce training time and variability. A trade-off is that the model is not able to predict higher prices.

3.3.1 Feature encoding

The data set had a variety of data types. Categorical features were encoded one-hot encoding technique. One-hot encoding transforms a nominal feature with n observations and d distinct values to d binary features that can take the values 0 and 1, indicating absence and presence respectively (Müller & Guido 2017). Table 2 illustrates the one-hot encoded features in the data set. The feature review_score was on the other hand manually encoded into binary values of (0,1) where 0 indicated a negative review and 1 a positive review. Negative reviews were considered to have a score of less than three and a score of higher than three if the review was positive. Feature encoding is a required step since support vector machine can only compute numerical values.

Table 2: Visual presentation of one-hot encoding.

payment_type		credit_card	debit_card	voucher	boleto
credit_card		1	0	0	0
debit_card	→	0	0	1	0
voucher		0	1	0	0
boleto		0	0	0	1

3.3.2 Feature selection

Feature selection is an important method that is used to reduce the number of non-informative input features which can degrade prediction accuracy. Selecting only relevant features can also prevent a model from overfitting and reduce computational cost.

First, intercorrelation detection was used to determine whether there were any relationships among the input features as this can cause misleading results. The selected method for the task was to use Pearson correlation coefficient also known as Pearson's r. The value of the coefficient *r* can range from -1 to +1. A value of -1 indicates that the relationship between the features has a perfect negative relationship whereas +1 is an indicator of a perfect positive relationship. If the value is 0 it means that there is no relationship among the features (Schober et al., 2018). Here, a threshold was set to cut off any intercorrelated features with a value higher than 0.50.

The second step was to use random forest to select features with the highest importance. The importance for each feature is determined by decreasing the impurity of variance. Table 3 shows the selected input features.

Table 3: Presentation of the selected input features.

freight_price	*product_description_length*	*product_name_length*	*product_weight_g*	*product_length_cm*	*product_height_cm.*

3.3.3 Feature scaling

All features except the encoded ones were standardized which scales numerical features to zero mean and unit variance. This is done because features can have different units and magnitude which means that features with higher magnitude can dominate over features with smaller magnitude. Ultimately, non-scaled features could be problematic especially when measuring the performance with RMSE because it gives more weight to larger errors.

3.4 Fitting the model

After feature scaling data was partitioned into 80% training data and 20% test data. The training set consisted of 13462 samples and the test set of 3411 samples. The last step was to implement a grid search with three-fold cross-validation to find the best hyperparameters and to fit the model along with the optimized parameters. Table 4 shows the unique settings for each parameter.

Table 4: A presentation of hyperparameters used in grid search.

Kernel	Epsilon	Gamma	C
RBF	0.001	0.01	0.001
	0.01	0.1	0.01
	0.05	1	0.1
	0.1	10	1
	1	100	10
	10		100

Chapter 4

Results & discussion

This chapter presents the results that were obtained from the methodological approach in chapter 3 along with a discussion. First a grid search over hyperparameters was carried out. The results from the grid search are presented in table 5. The parameter values were then used during the training of the model.

Table 5: The table shows the best hyperparameter through grid search.

Kernel	Epsilon	Gamma	C
RBF	0.1	100	100

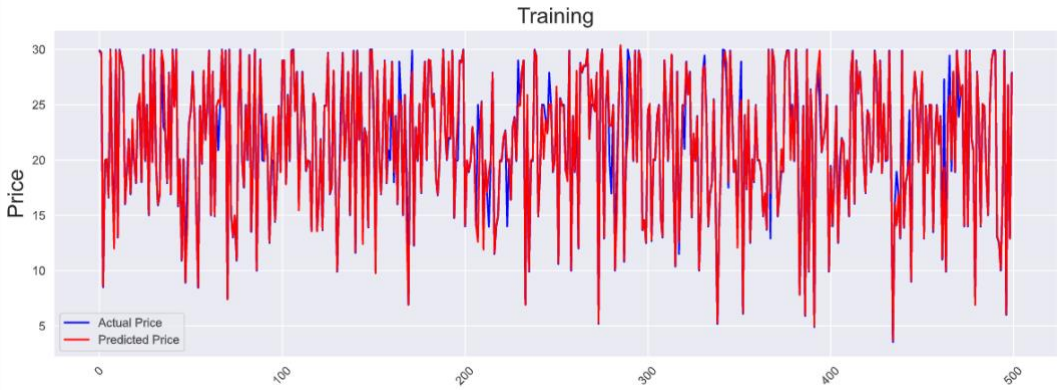

Figure 4: Predictions on training data.

In figure 4 it can be observed that the model fits the training data well and is able to capture the underlying patterns in the data. Predictions on the training data gave an MAE of 0.533 and RMSE of 1.484.

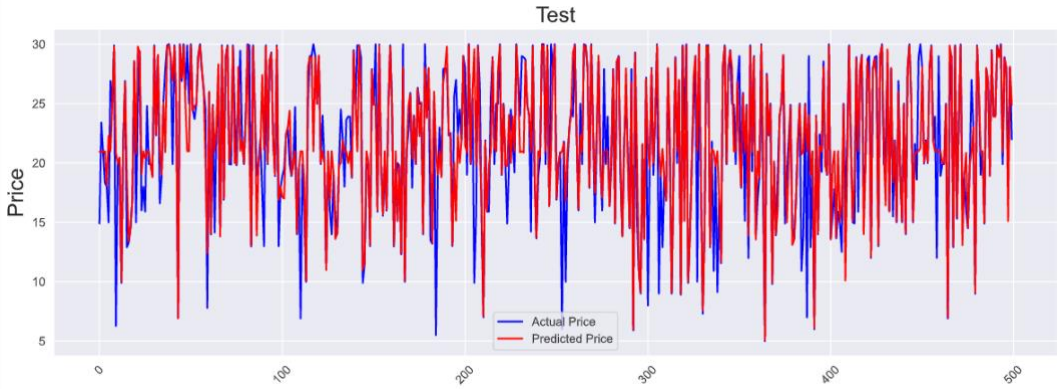

Figure 5: Predictions on unseen data.

Meanwhile, predictions on the test set presented an MAE of 1.865 and RMSE 3.604. Figure 5 shows how the model is less consistent in its predictions. This can be seen from the blue peaks in the graph. The MAE score implies that the distance between the predicted price and the actual price is on average off with approximately 1.865 units. Similarly, the RMSE score implies the squared differences.

Although obtained results show good accuracy as seen in the results, the test scores are not close to the scores obtained on the training set which indicate that there might be a case of overfitting. A likely reason for this might be that selected hyperparameters from running the grid search are not optimal. It is also not clear what hyperparameters are to be tested. This is also stated in the study conducted by (Siwers & Dahlén, 2017) who presents a similar explanation of their results. The preprocessing steps might also have been insufficient with respect to the difficulties of the data. This means that the model was likely trying to learn from the underlying relationship between the input and output values along with unidentified noise in training data.

A modification to the used methodological approach would be to spend more time identifying and removing any unexpected anomalies from the data set. From viewing the selected input features, it could also be discussed if they truly affect prices in reality. Whether the errors are within an acceptable range is on the other hand difficult to determine due to lack of extensive research of machine learning and price prediction of products. Being off with 1.865 units from the actual price could significantly impact business profits negatively. On the other hand, considering the price of a specific product over time-horizons could have lowered the errors which was demonstrated by (Bakir et al., 2018) as prices commonly are affected by seasonality. This also means that the suggested model would be more useful in the real-world. However, this would only make sense when price history of products is available. Many organizations don't reveal their price history and extensive data for competitive purposes. Therefore, the results obtained in this study are limited and cannot be compared to the results obtained from other support vector machines in the reviewed literature. The observed results in other studies are also measured with different metrics which cannot be compared with the results in this study.

Chapter 5

Conclusion & future work

In this study real-world e-commerce data and support vector machine is used to predict product prices. The study begins by conducting a literature review to summarize previous work and theoretical concepts gathered from books, scientific papers, and open articles. Then, an experiment is undertaken to verify the research question. The experiment begins by collecting real e-commerce data which is then preprocessed. After that, a grid search over optimal hyperparameters with a three-fold cross-validation is used. After that, a support vector machine with a RBF-kernel is trained with the preprocessed data along with the optimal hyperparameters obtained from the grid search. From the results in this study, it can be concluded that support vector machine holds potential in accurately predicting the price of products in a real-world application. The proposed model is also more flexible than simple parametric models such as linear regression since it makes less assumption about data and can solve nonlinear problems. It is however evident that there are limitations with respect to the data and the approach.

This gives room for future work with support vector machine and price prediction of products. Consequently, demonstrated techniques and approaches in this study can serve as a guideline to using support vector machine to predict the price of products on other data sets. Further suggestions to expand the scientific research in this area would be to try neural network or other nonparametric machine learning models on the same data set and compare the results. For example, recurrent neural network with long short-term memory have shown to outperform support vector machine in certain situations. It could also be discussed whether the obtained results could be of interest in areas other than the space of e-commerce. (Siwers & Dahlén, 2017) claim that businesses that incorporate data into their decision-making can be more profitable. The approach of price prediction could be relevant in the hotel or airline industry to help businesses optimize their room or flight ticket prices.

References

Bakir, H., Chniti, G., & Zaher, H. (2018). E-Commerce Price Forecasting Using LSTM Neural Networks. *International Journal of Machine Learning and Computing,* 8(2), 169–174. https://www.researchgate.net/publication/325686014_E-Commerce_Price_Forecasting_Using_LSTM_Neural_Networks

Gupta, R., & Pathak, C. (2014). A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing. *Procedia Computer Science,* 36, 599–605. https://www.researchgate.net/publication/275541641_A_Machine_Learning_Framework_for_Predicting_Purchase_by_Online_Customers_based_on_Dynamic_Pricing

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88

Liang, X., Zhang, H., Xiao, J., & Chen, Y. (2009). Improving option price forecasts with neural networks and support vector regressions. *Neurocomputing,* 72(13–15), 3055–3065. https://www.sciencedirect.com/science/article/pii/S092523120900109X

Yu, X., Qi, Z., & Zhao, Y. (2013). Support Vector Regression for Newspaper/Magazine Sales Forecasting. *Procedia Computer Science*, 17, 1055–1062 https://www.sciencedirect.com/science/article/pii/S1877050913002676

Bakshi, C. (2020, June 4). *Support Vector Regression - The Startup*. Medium. https://medium.com/swlh/support-vector-regression-explained-for-beginners-2a8d14ba6e5d

Kedia, S., Jain, S., & Sharma, A. (2020). *Price Optimization in Fashion E-commerce.* KDD 2020 Virtual Conference, California. https://www.researchgate.net/publication/342886643_Price_Optimization_in_Fashion_E-commerce

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.) [E-book]. O'Reilly Media.

Abellera, R., & Bulusu, L. (2017). *Oracle Business Intelligence with Machine Learning: Artificial Intelligence Techniques in OBIEE for Actionable BI* (1st ed.) [E-book]. Apress.

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (1st ed.). Apress. https://www.researchgate.net/publication/277299933_Efficient_Learning_Machines_Theories_Concepts_and_Applications_for_Engineers_and_System_Designers

Sharda, R., Turban, E., Delen, D., Efraim Turban, Dursun Delen, Aronson, J. E., Liang, T. P., & King, D. (2014). *Business Intelligence and Analytics* (Global Edition). Pearson.

Oates, B. J. (2006). Researching Information Systems and Computing (First ed.). SAGE Publications Ltd.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data* (1st ed.) [E-book]. O'Reilly Media.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists* (1st ed.) [E-book]. O'Reilly Media.

Johansson, D. (2020). *Price prediction of vinyl records using machine learning algorithms*. [Bachelor's thesis, Linneaus University]. DIVA. http://www.diva-portal.org/smash/get/diva2:1443317/FULLTEXT01.pdf

Unctad. (2021, March 15). *How COVID-19 triggered the digital and e-commerce turning point.* https://unctad.org/news/how-covid-19-triggered-digital-and-e-commerce-turning-point

Ihre, A., Engström, I. (2019). *Predicting house prices with machine learning methods.* [Bachelor's thesis, KTH]. DiVA. http://kth.diva-portal.org/smash/get/diva2:1354741/FULLTEXT01.pdf

Siwers, R., Dahlén, C. (2017). *Predicting sales in a food store department using machine learning.* [Bachelor's thesis, KTH]. DiVA. https://www.diva-portal.org/smash/get/diva2:1108597/FULLTEXT01.pdf

Yadav, A. (2018, October 20*). SUPPORT VECTOR MACHINES(SVM) - Towards Data Science. Medium.* https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

Claesen, M., & de Moo, B. (2015). Hyperparameter Search in Machine Learning. https://www.researchgate.net/publication/272195620_Hyperparameter_Search_in_Machine_Learning

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia, 126(5), 1763–1768. https://www.researchgate.net/publication/323388613_Correlation_Coefficients_Appropriate_Use_and_Interpretation

Elraffah, Z. (2021, March 9). Price optimization with machine learning: what every retailer should know. 7Learnings. https://7learnings.com/blog/price-optimization-with-machine-learning-what-every-retailer-should-know/

Kaggle. (n.d). *E-Commerce Data* [Dataset]. Kaggle. https://www.kaggle.com/carrie1/ecommerce-data

Appendix A

Tools

Following definitions are obtained from (Müller & Guido 2017).

Python
A general-purpose programming language suitable for data science.

Jupyter Notebook
A web-based application for running Python.

Pandas
A data processing library used for data manipulation and data analysis in Python.

Numpy
Similar to Pandas library but with added scientific computing capabilities.

Matplotlib
A data visualization library for creating interactive plots.

Scikit-Learn
A machine learning library that is used in Python.

Appendix B

Data

Table 6: Explanation of features of the original data set (Kaggle n.d).

Feature	Description
order_id	unique order id
customer_id	key to the orders data set
order_status	order status
order_purchase_timestamp	purchase date
order_approved_at	payment approval date
order_delivered_carrier_date	timestamp when handled to carrier
order_delivered_customer_date	actual delivery date
order_estimated_delivery_date	estimated delivery date to customer
payment_sequential	number of different payment methods chosen by customer
payment_type	method of payment
payment_installments	number of installments
payment_value	transaction value
customer_unique_id	unique id of a customer
customer_zip_code_prefix	customer zip code
customer_city	customer city name
customer_state	customer state
order_item_id	sequential number identifying number of items included in the same order
product_id	id of product
seller_id	seller unique identifier
shipping_limit_date	limit date for handling the order to logistic partner
price	item price in brazilian real (R$)
freight_value	freight value
product_category_name	category in portuguese
product_name_lenght	character length of product name
product_description_lenght	character length of product description
product_photos_qty	number photos
product_weight_g	product weight in grams
product_length_cm	product length in centimeters
product_height_cm	product height in centimeters
product_width_cm	product width in centimeters
product_category_name_english	category name english
review_id	review id
review_score	review score ranging from 1 to 5
review_comment_title	review title in portuegese
review_comment_message	review message in portuegese

review_creation_date	date when satisfaction survey was sent to the customer
review_answer_timestamp	satisfaction survey answer timestamp