E-commerce Price Suggestion Algorithm – A Machine Learning Application

Linh Nguyen Tran Quang
Student, School of Industrial Engineering and Management,
International University -Vietnam National University,
Ho Chi Minh City, Vietnam
ntquanglinh1999@gmail.com

Uyen Ngo Thi Thao, Anh Duong Vo Nhi, Hoang Nguyen Thi Phuong
Lecturer, School of Industrial Engineering and Management,
International University -Vietnam National University,
Ho Chi Minh City, Vietnam
nttuyen@hcmiu.edu.vn, dvnanh@hcmiu.edu.vn,
phuonghoangnguyenthi11@gmail.com

Abstract

E-commerce is a key move that has altered the way businesses are conducted, particularly in the retail industry. In the e-commerce business, pricing is one of the most significant aspects in determining profitability, and closely linked to the company's sales. The rise of online markets has necessitated the creation of a Machine Learning tool for pricing suggestions. The aim of this study is to select a machine learning model to create a price suggestion tool for Ecommerce enterprises. Three machine learning algorithms – Linear Regression, Random Forest and LightGBM – are tested on a dataset of an Ecommerce enterprise to indicate the performance of models when using in a dataset with several features and millions of rows. The study also processes to make improvement the output of models, including parameters tuning, feature selection or remove outstanding values. The result shows that Light GBM after Grid search CV process outperforms in terms of both prediction error and processing time.

Keywords
E-commerce, LightGBM, Machine learning, Price suggestion, and random forest.

1. Introduction

E-commerce is a significant step that has changed the way of doing business, especially in retail market, plays an important role in the development of the national economy in the 21st century. Simultaneously with the development of the E-commerce industry, the massive growth of new terms such as Machine Learning (ML), Data Science (DS), Deep Learning (DL), Artificial Intelligence (AI) have also been paid attention due to their impact on the retailing market (Chandrashekhara et al., 2019; Jia et al., 2013; Narayana et al., 2021; Pundir et al., 2020). Akter and Wamba (2016) believe that data plays an important role in E-commerce and all business decisions. Furthermore, the enhancement in data availability as well as processing speed allows E-commerce enterprise to deal with complex problems requiring a large enough dataset by using Machine Learning applications. Machine Learning has been currently applied in some case of E-commerce such as Last Mile delivery, Synchronized system and made predictions with strong accuracy, responded and interacted with inconsistent market demand or reduced operations costs (Hurtado et al., 2019). Pricing is also an important term not only in E-commerce but also all business market in the act of determining, specifying the value of product before selling them to customers to achieve profit maximization, competitors' matching price, market skimming or even long-term survival. Thus, predicting price is a crucial part in e-commerce and other market sections for planning and strategy development. The combination of the price prediction with the application of machine learning is becoming an increasing trend in both researching and practical management expertise.

By introducing the suitable method for price prediction, it is a good opportunity for E-commerce platforms to automatically set the reasonable prices on massive products, reducing mistake of doing manually. The main objective of this article is to apply, analyze and evaluate several machine learning algorithms for predicting the price of multiple products exhibiting on the Ecommerce site. While establishing models, parameters are consecutively changed to find the best sets in each algorithm. Boosting method has also been used to enhance the accuracy of models, minimize processing time of the models and limit overfitting or underfit situation. Through this study, E-commerce enterprises can generate an automatically price suggestion tools, which eases sellers and E-commerce platforms come to an agreement in pricing of products.

2. Literature review

Machine Learning is the scientific study of algorithms and mathematical models that computers are willing to learn and effectively perform a selected task, count on patterns within the data and illation instead. It is closely associated with statistics analysis, using computer algorithms to create predictions for new instances by learning patterns of datasets. The study of mathematical optimization provides theoretical foundation and practical application to the sector of machine learning (Mohri et al., 2018). Popularly, machine learning algorithms are classified into 4 main classes: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Géron, 2019).

Pricing has become one of the most significant research fields having been studied in combination with machine learning tools, especially E-commerce industry. In a research article, Gupta and Pathak (2014) apply machine learning in predicting dynamic pricing. The models are trained in a large dataset in Ecommerce with specific features and divided them into some main groups of merchandises and groups of customers. With the application of K-means clustering and Logistic Regression, the result shows that a reasonable price for each group of customers led to an improvement in terms of revenue.

Fathalla et al. (2020) introduced a supervised machine learning to create model to predict prices of second-hand commodities. In this work, Time Series method and Linear Regression are used to map the features of the products with their price. The results show that the model can be applied to the datasets with different products while providing acceptable scores.

Tziridis et al. (2017) identifies the price of airfare of a European airline by using 7 Machine Learning models. The models were respectively tried in many combinations of features to find out which features were played an important role that affect the results. The results indicated Random Forest is the most trustworthy model as the most stable model supported by its average accuracy. Moreover, feature selection plays a key role before applying Machine Learning models to reducing processing time as well as increasing the accuracy of the price.

The most recent research in price prediction that using Machine Learning to predict prices for a C2C Ecommerce company in Asia is Chada (2019), in which several Machine Learning models were proposed to forecast a price of used products with different sets of attributes. Feature extraction was used to deal with unstructured parameters that transform data into suitable forms so that models could process. The results pointed out that product features played an important role in predict price of product. Some modern Machine Learning algorithms such as LightGBM algorithm or Ridge Regression had a huge benefit to carry out complex, numerous datasets while ensuring the accuracy score at the high levels. Despite of its size, pictures could be a useful factor to enhance accuracy in the future. Following the articles in literature, the study of Chada (2019) could be considered as the most related to the problems that this study needs to deal with. The paper concerns about the unstructured data and how to convert it to the form that can be read by the models. The application of two new Machine Learning algorithms, including LightGBM and Ridge Regression, was also deeply investigated due to their performance with large datasets.

3. Methodology

There are numerous numbers of Machine Learning algorithms for predicting prices being used in the literature, including Multiple Linear Regression (MLR), Support Vector Machines (SVM), LightGBM (LGBM), K-Nearest Neighbors (KNN), Decisions Trees (DT), Random Forest Regression (RFR). After a sufficient comprehension of the advantages and disadvantages as well as their operations, it is necessary to consider few algorithms to address the problem of price prediction. The accuracy of algorithms in a predictive problem, processing speed and dealing with overfitting are three evaluation metrics that need to be considered to select the reasonable methods. Multiple Linear Regression, LightGBM and Random Forest are chosen to apply for a case study of predicting prices for a Japanese E-commerce company. To sum up, this study refers to develop a regression model by using both traditional (Linear Regression and Random Forest) and modern (LightGBM) Machine Learning algorithms with the goal to suggest a suitable price for a C2C Ecommerce platform. Through this study, the performance between traditional and modern technique as well as the effectiveness of those methods could be evaluated in a real case.

4. Model development and Solution generation

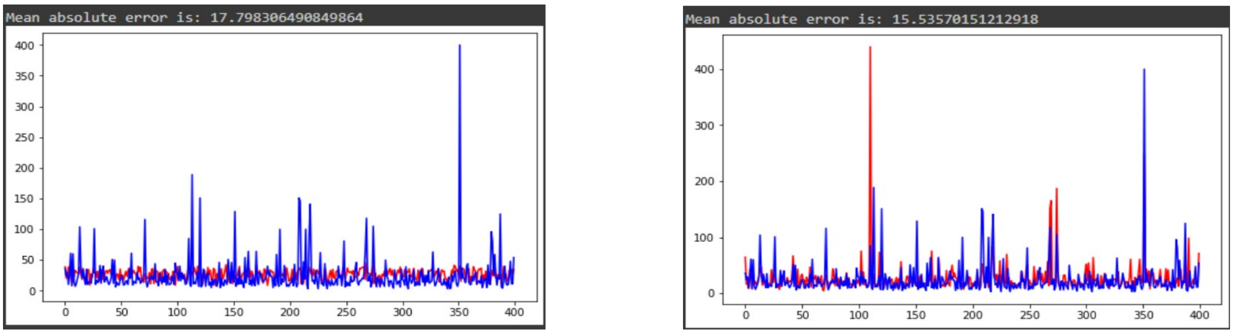
4.1 Data collection, preprocessing, and exploration

The planned models were trained in a public real dataset from Mercari - a Japanese C2C Ecommerce company who wants to offer a pricing suggestion tool for sellers, but it is hard due to the complex of data that sellers upload on Mercari marketplace. This dataset is a part of Mercari competition that they challenged participant to create a suitable Machine Learning algorithm for pricing recommendation, containing 7 features and almost 1.5 million observations. Founded in 2013, Mercari, Inc. is a Japanese company that operates one of the most popular C2C marketplace in Japanese market. From the beginning of Mercari, their objective is to create a popular, easy-access platform that users can sell and buy their own products without limitations of time, space, and prices. The dataset provides several information that sellers input, including products description, category name, brand name and condition of products. The dataset is collected primary in 3 main categories: Fashion, electrical devices, and cosmetics. Those products are extremely sensitive with price and customers carefully consider before making buying decisions.

An overview to the dataset and each feature leads to a specific number of steps which are required for pre-processing. According to the overall report, there were some duplications of rows that should be dropped. Some of unused columns are also removed before executing preprocessing data. After that, creating some new columns, filling not available values, removing useless rows, exploring deeply the dataset are needed to be fulfilled to provide a suitable dataset for training and testing. As dataset is quite large, data preprocessing before training and testing plays a key role in the result. New features are created from original feature, not available values are replaced, and missing values are dealt with. To proceed categorical data, Label encoding and One-hot encoding are employed (Yu et al., 2022). After preparation, the dataset is divided into training set with 80% of the dataset and 20% for the test set. Due to limited in space, data preprocessing and data exploratory analysis are not described in detail in this paper. After running algorithms in Python, mean absolute error (MAE) is the primary metric to evaluate the model performance. The strength of this metric is the unaffected by the direction of the value and the bias reduction when dealing with high values (Willmott and Matsuura, 2005).

4.2 Basic model generation and initial results

In this section, 3 models (Linear regression, Random Forest, LightGBM) are run with default setting to obtain basic results as foundation for further improvement. It can be seen from Figure 1 that when using Linear Regression to predict, actual price values have a wide range from 0 to more than 400. The red line indicates the prediction of the model while the blue one shows the real values. Linear Regression works well for the price under 50, when the price is bigger than 50, there is a big gap between the real and predicted numbers. The intercept of this model is 32.67 and all of features in hand are used. Removing the intercept, adjusting the number of features used or modifying the range of price are some reasonable solutions to be considered in the following part. The Random Forest Regressor takes the longest time to proceed with the default parameters. It is a weakness of this algorithm that takes too much time, and this is also a barrier for learner in case of tuning parameters. As can be seen from the chart in Figure 1, the prediction from Random Forest seems to be more exact than the Linear Regression. The distance between actual values and forecast values, especially with the big value, is significantly reduced which indicate that this model has an ability to make prediction even with the outstanding value. This is again confirmed by detailed result in Table 1, in which Linear Regression has MAE of 17.79 while Random Forest has only 15.53. Also being shown in this Table, the result from Light GBM is quite impressive that it just needs only nearly one minutes to finish both training and testing.

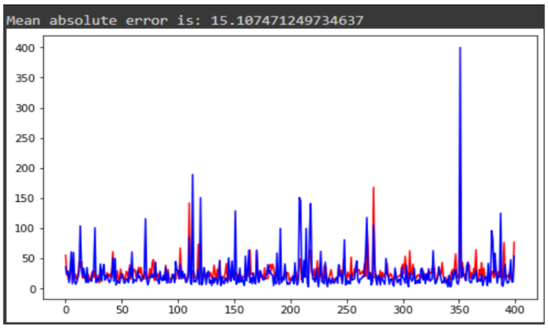


```
Mean absolute error is: 17.798306490849864
```

(a)

```
Mean absolute error is: 15.53570151212918
```

(b)



```
Mean absolute error is: 15.107471249734637
```

(c)

Figure 1. Linear Regression (a), Random Forest Regressor (b) and LightGBM (c) initial results

Table 1. Initial result summary

	Actual Price	**Linear Regression**	**Random Forest**	**LightGBM**
Count	296482	296482	296482	296482
Mean	26.84	26.72	26.97	26.76
Std	38.93	8.02	24.37	17.26
Min	0	3.21	3	5.06
Q1 (25%)	10	20.72	15.17	17.63
Q2 (50%)	17	26.92	20.76	22.46
Q3 (75%)	29	32.81	30.77	29.72
Max	1625	61.76	1211.12	375.91
Processing time		10s	470s	15s
Mean Absolute Error		17.79	15.53	15.15

As can be seen from the Table 1, these are statistics specifications of predicted values in comparison with the target variables, processing time and error of all algorithms. There are totals of 296482 values in the test set to make evaluation to the outcome of these methods. Mean values of all algorithms are seemed to be not very different (around 26.8). In terms of processing time, Linear Regression outperforms the other two models, but error value is the highest. Meanwhile, LightGBM is a good choice with short processing time (15s) and low error (15.15), slightly lower than error of the Random Forest model (15.53).

4.3 Improvement results

As mentioned above, the initial result is based on the default setting of each algorithm. Those settings need to be modified along with adjusting the dataset to reach the maximum possible performance of the model. Each model eliminates some features and uses some specific ones or wipes out some of range of price to enhance this problem. Cross-validation is a technique for evaluating models that estimate their competence in a small sample sequentially (A Ramezan et al., 2019; Refaeilzadeh et al., 2009; Schaffer, 1993). According to initial results of three models, predicted values are quite close to the actual values when the actual values are under 3-digits. The range of price is widely from 0 to nearly 2000 but most of values are under 100. Although the number of 3-digits values is not numerous if compared with the whole dataset, the deviation between predicted and actual range can seriously affect the performance of models. Based on reasons above, removing some range of values is necessary. There are many ways to remove outstanding values and interquartile range is a statistical measurement to determine outstanding values, or outliers (Walfish, 2006). The interquartile range (IQR) is defined by take Q3 – Q1 where Q3 and Q1 are 75% and 25% alternatively. The data will be eliminated if they are bigger than Q3+1.5IQR or smaller than Q1-1.5IQR. After calculating, value is considered as outlier if they are bigger than 57.5 or smaller than -18.5. Because the minimum value is 0 so the dataset just only has upper bound fence. In addition, 0 need to be removed if the customer navigates the product with no value or this is a present; thus, it is no need to give them a suggestion about pricing. Aggregating all conditions, there are approximately 120,000 rows which are cut out of the dataset (about 8% of dataset). After removing outliers, the distribution of dataset is illustrated in Figure 2.

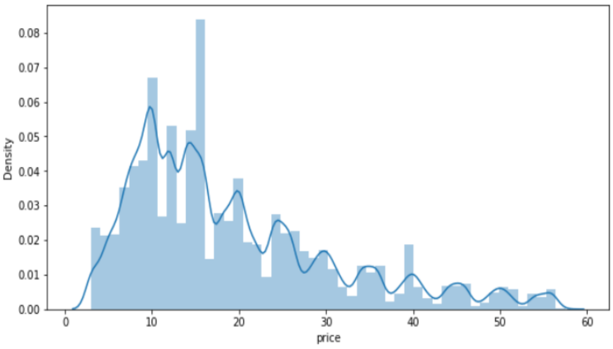
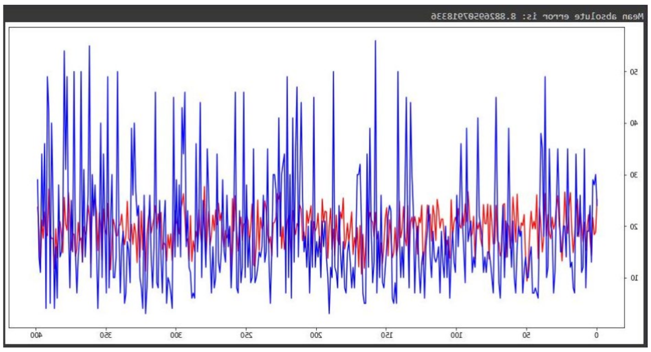


Figure 2. Dataset after removing outliers



```
Mean absolute error is: 8.88269507918336
```

Figure 3. Linear Regression improved result

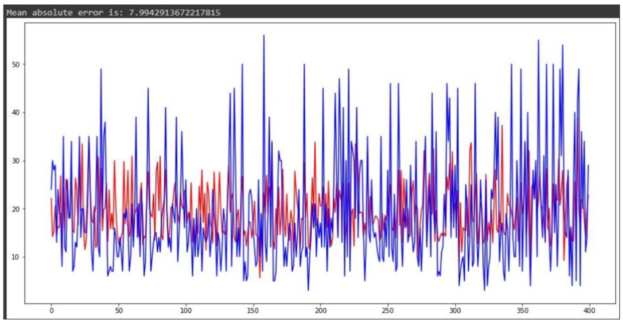
Linear Regression has one option to change parameter that whether the intercept is used or not. The change does not bring any changes regarding to the result. In addition, the dataset after removing outstanding values brings a huge benefit that the error reduces significantly, as being shown in Figure 3.

For Random Forest model, GridsearchCV technique is applied to find the optimal parameters (Grgić et al., 2021; Paper and Paper, 2020; Ramadhan et al., 2017). Table 2 shows the result of Random Forest optimal parameters suggested by GridsearchCV. There is total 48 combinations that this method will run to find a set of optimal parameters. This process takes more than two hours to finish all combinations. The new set of parameters brings a dramatic result when applying to the new dataset. As can be seen in Figure 4, the predicted value is very close with the test value, especially with prices in the range from 10 to 35. However, with the price is bigger than 35, it shows that the model cannot reach, and that product is considered as overrated.

Table 2. Parameters before and after using GridsearchCV (Random Forest)

	Initial value (before Gridsearch-CV)	**Optimal value (after Gridsearch-CV)**
Max_depth	None	10
N_estimators	100	200
Min_samples_leaf	1	1
Min_samples_split	2	2

Both LightGBM and Random Forest have the same father - the Decision Tree algorithm (Ke et al., 2017; Minastireanu and Mesnita, 2019). However, LightGBM is one of the best Gradient Boosting methods in terms of the time-consuming issue and remains a reasonable result in comparison with others. Because of having the same logic, LightGBM also has some similar parameters like Random Forest. After considering parameters as result in Table 3, it comes up with the setup of LightGBM model with 243 combinations during this process. Because of the superiority in time-consumption, it brings a huge confidence to try in a lot of combinations.



```
Mean absolute error is: 7.9942913672217815
```

Figure 4. Random Forest result after using new set of parameters

Table 3. Parameters before and after using GridsearchCV (LightGBM)

	Initial value (before GridsearchCV)	**Optimal value (after GridsearchCV)**
Max_depth	-1	-1
Num_leaves	31	50
Min_child_samples	20	50
N_estimators	100	800

The result in Table 4 indicates that the improved LightGBM has the best performance in comparison with other algorithms. Overall, the predicted values do not show big difference compared with the former one although error is lower (7.63 compared with 7.99). Table 4 shows the summary of model outputs from 3 algorithms after refining dataset and tuning parameters. As can be seen from the table, there is a drop of the number of rows and the maximum of the price because of the elimination of some outstanding values. Although the number of eliminated rows is a small proportion of the whole dataset, that decision gives a better solution all of errors are nearly equal one-half of the initial results.

Table 4. Result after refining dataset and parameters tuning

	Actual Price	**Linear Regres-sion**	**Random Forest**	**Light-GBM**
Count	272438	272438	272438	272438
Mean	19.22	19.20	19.19	19.21
Std	11.88	3.62	5.56	6.34
Min	3	7.7	4.71	3.49
Q1 (25%)	10	16.56	4.71	14.71
Q2 (50%)	16	19.18	18.77	18.28
Q3 (75%)	25	21.97	22.35	22.84
Max	57	30.32	47.43	50.65
Processing time		2s	490s	60s
Mean Absolute Error		8.88	7.99	7.63

Moreover, algorithms have been conducted to tun parameters to choose the reasonable set of parameters, which enhances the performance of these predictive models. Appropriate set of parameters optimize the efficiency of models in a various training/testing dataset. However, the processing time cannot be enhanced because it takes time to print out a reasonable solution. LightGBM is still a leader of all algorithms that all statistical indexes are closed to the ones in the price column. Mean values of all algorithms are not seemed to be very different (around 19.2).

4.4 Feature re-selection

Feature selection is an approach of cutting the number of predictors to decrease computational effort of a predictive model, thus improving the model's performance (Kumar and Minz, 2014; Li et al., 2017). Feature selection can be applied on Random Forest and LightGBM. There are total of 19 features, the study would conduct experiments at 5, 8, 10 and 15 features ascending in feature importance. Both models remain the same sets of parameters that are based on the optimal parameter sets resulted from the GridsearchCV.

4.4.1 Random Forest

Before re-selection feature, a list of features will be printed out and their ranks are based on how importance that each feature contributes or affects to the predicted value. Table 5 shows importance of all features in percentage according to ascending order. The most important set of features will be selected to the model (Figure 5).

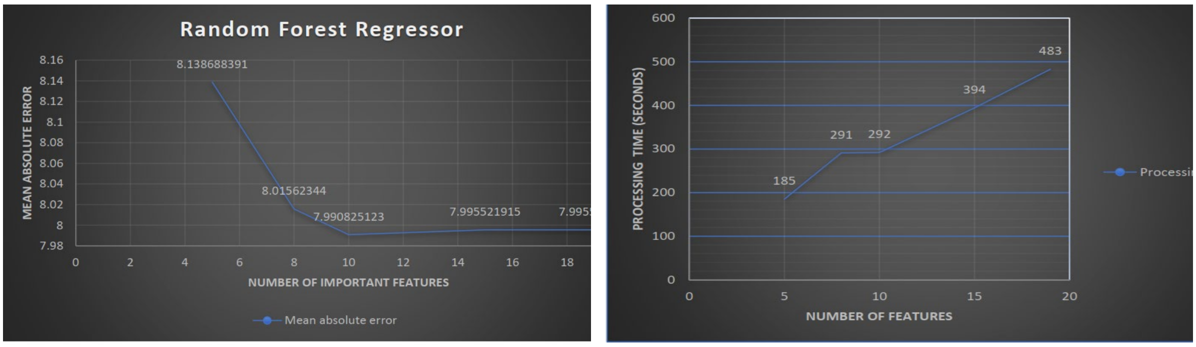
Table 5. Feature importance in Random Forest

Rank	**Feature**	**Importance**
1	subcat_1	0.232951
2	subcat_2	0.218938
3	shipping	0.177124

4	item_condition_id	0.104164
5	famous brand	0.069633
6	women	0.06840
7	description_length	0.038963
8	no brand	0.025718
9	men	0.016337
10	kids	0.015839
11	beauty	0.011528
12	electronics	0.008530
13	unpopular brand	0.006089
14	handmade	0.003857
15	home	0.001184
16	vintage and collectibles	0.000298
17	other	0.000280
18	sport and outdoors	0.000160
19	no label	0.000008
Total		1

It is apparent that the training and testing process will be processed faster than when using the original dataset because some of columns will be removed if there are useless. As a result, the amount of data needed to be processed will be reduced. If the error remains unchanged or slightly increases but the processing time is enhanced significantly, learners should consider to permanently eliminate some unnecessary features. Figure 5(a) and 5(b) show performance and processing time of Random Forest model given the most 5, 8, 10, 15 and 19 important features.

As can be seen from the Figure 5(a), there is a downtrend of MAE when the number of features increases to the maximum level. However, when looking at the number of features from 10 to 19, the difference between them, which is not apparent, could be ignored. There is an opposite trend in the chart in Figure 5 (b) that the more information means that the more processing time needed to achieve the result. The trend significantly increases from 10 to 19 features and reached the peak at 19 features (483 seconds). Combinations with the view from the first chart, 10 features are the ideal features which indicates that it does not need all features to get the reasonable result. Using 10 features can provide the best solutions while the processing time can reduce up to 40% compared with using all features (Figure 5).



(a)

(b)

Figure 5. Random Forest model performance and processing time with different number of features

4.4.2 LightGBM

Although processing time of this method is good enough compared with the former one, reselection feature is also important that learners can have more chances and abilities to adjust more parameter to find out which set of parameters is the best one for their problem. LightGBM is also applied the same tactic of adjusting number of features. Model is run with different numbers of most important features, as being shown in Table 6 and Figure 6.

Table 6. Performance of model after feature re-selection (LightGBM)

Number of features	**Mean absolute error**
5	7.812955440705378

8	7.651544494097297
10	7.636582839129962
15	7.6389543387925904

As can be analyzed from the table, prediction error tends to reduce gradually when being added more features into the algorithm, similar to the case of Random Forest abovementioned. Since more than 10 features cannot help to increase its accuracy, 10 feature is the ideal value that the error is slightly lower than using all features although the difference is not clear (7.6365 compared with 7.6367), but it supposes to reduce computing time. The processing time of this algorithm is not mentioned in case of using all features with only 60 seconds. Therefore, if less than the total of features is used, the processing time must be shorter than before, but the difference might be only some few seconds.

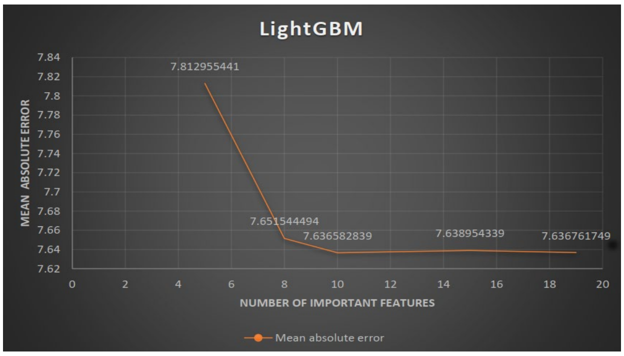


Figure 6. Model performance with different number of features (LightGBM)

4.5 PCA for dimensionality reduction

Principal Component Analysis (PCA) is a technique to reduce data dimensionally, thus transforming a large data set into a smaller one while preserving most of the information of the original data set. There is a loss of information as cutting some variables, PCA naturally lowers model accuracy. However, the main point of dimensionality reduction is to tradeoff between accuracy and simplicity. Since it is simpler to examine the smaller amount of data, machine learning algorithms can proceed data faster without having to deal with redundant factors (Bro and Smilde, 2014). After analyzing, it is observed that 3 is the ideal number of components that contains most of information of the former dataset (approximately 100%). The used algorithms are improved models after tuning using GridsearchCV (section 4.3).

After running 3 algorithms, results are shown in Table 7. It is shown that mean value of all algorithms is quite close to the actual value, which is around 19.22. Although the amount of data is significantly reduced due to applying PCA that only keeps 3 features, the processing time of the methods do not improve as expected in comparison with the case of full dataset in section 4.3. Same thing happens for error, and PCA even causes the algorithms to have the worst accuracy compared with other methods. This probably can be explained that PCA cuts down some information of the dataset that are extremely important to the model (Table 7).

Table 7. Results of tunned model before and after PCA

	Tunned model before PCA			Tunned model after PCA		
	Linear Regression	**Random Forest**	**LightGBM**	**Linear Regression**	**Random Forest**	**LightGBM**
Mean	19.207	19.199	19.210	19.204	19.213	19.210
Processing Time	2s	490s	60s	1s	590s	50s
Mean Absolute Error	8.88	7.99	7.63	9.389	8.585	8.247

4.6 Shipping and without shipping

According to feature importance in previous section, Shipping ID is considered as one of the most important features. The special of this feature is that it only has 2 values are 0 and 1 representing which parties (customer or seller) will pay for the transportation cost. If the Shipping ID value is 0, seller is the person in charge of this cost and vice versa. In addition, a product's price in which shipping cost is paid by the customer is smaller than the one is paid by the seller. Since the product's price is higher when seller takes responsibility for shipping cost, the product's price is not a real value of it because seller must add an extra fee to cover the shipping cost. As a result, the dataset is divided into 2 parts, according to shipping ID status, 0 and 1. For each half of the data set, all algorithms work well with dramatically reduced processing time, which is understandable since data amount is reduced. The results are shown in Table 8.

Table 8. Model results in 2 cases of shipping ID

Shipping ID = 1				
	Actual Price	**Linear Regression**	**Random Forest**	**Light-GBM**
Count	124584	124584	124584	124584
Mean	16.63	16.59	16.60	16.60
Std	11.45	3.10	5.63	6.15
Min	3	7.38	4.23	0.38
Q1 (25%)	8	14.08	13.27	12.27
Q2 (50%)	13	16.59	14.81	15.26
Q3 (75%)	22	18.92	18.67	19.51
Max	57	25.74	48.33	44.18
Processing time		1s	170s	24s
Mean Absolute Error		8.51	7.41	7.13
Shipping ID = 0				
	Actual Price	**Linear Regression**	**Random Forest**	**Light-GBM**
Count	147855	147855	147855	147855
Mean	21.42	21.41	21.43	21.42
Std	11.84	2.76	4.84	5.57
Min	5.5	11.36	11.14	6.55
Q1 (25%)	12	19.46	17.97	17.27
Q2 (50%)	18	21.55	20.35	20.34
Q3 (75%)	28	23.32	24.47	24.49
Max	57.5	34.17	50.05	50.00
Processing time		1s	225s	27s
Mean Absolute Error		9.14	8.37	8.09

As can be seen from the table, when splitting the dataset according to the shipping status, there are two obvious results. With shipping status equals to 1, predicted mean values of price are smaller (16.5 compared with 21.4), which is reasonable because without including shipping fee, price should be cheaper. Moreover, there is an opposite view of mean absolute error in two tables. With shipping status equals to 1, mean absolute error seem to be lower than the ones in shipping status of 0 with all algorithms, price prediction algorithms work better when no shipping fee included, in other words. LightGBM is still a leader that the performance of this algorithm is always at the best level with all method or experiments. The prediction error in the case of shipping cost included (shipping ID 0) is higher than the case of no shipping cost included (shipping ID is 1) may be because there is a wide range of shipping costs due to wide range of different merchandise segments, this may confuse the algorithms to detect the real prices of products, leading to high error in price prediction.

5. Sensitivity analysis

Sensitivity analysis, which looks at the effect of each feature to the prediction of the model, is a simple and effective approach to comprehend a machine learning model. In order to measure feature sensitivity, the value of a feature is modified or tried to examine the model's output after ignoring it in some way while leaving all other characteristics unchanged (Lenhart et al., 2002; Tortorelli and Michaleris, 1994). If altering the feature value substantially changes the model's outcome, this feature has a significant influence on the prediction. In this study, the main purpose is to assume the absence of the feature in the model.

It may accomplish that in models like neural networks by inserting zero. In additions, the possible way is to change the data by introducing it into the dataset using the mean for numerical features, a new class for categorical features, the highest probability value, or any other technique (Figure 7).

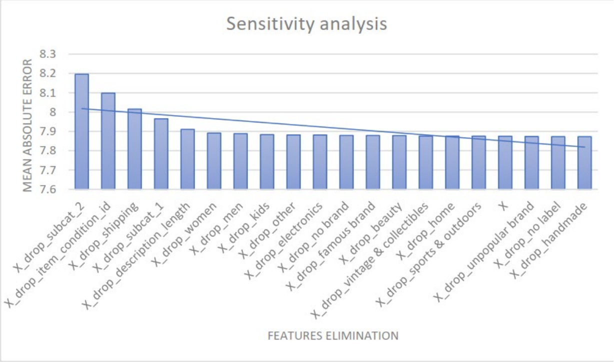


Figure 7. Sensitivity analysis result

As can be seen from the bar chart in Figure 7, trying to eliminate each of feature (exclude X) and then training/testing on the best algorithm – LightGBM before going through this experiment. There are top 5 features from the beginning that affect seriously to the results, while the others could be considered to eliminate do not have clearly affect to the result that when training/testing. From the fifth to the end, the result remains unchanged, and X is the original dataset that is a benchmark to make comparison with other cases. In conclusion, the model would perform better when some features are dropped, which are Subcat 1 and 2, Item condition, Shipping and Description length. Other than that, the performance remains stable.

5 . Conclusion

Price prediction has shown to be difficult for online marketplaces. Sellers are frequently overconfident, labeling items with selling prices that are far higher than they should be. The idea is to automatically determine a price for a product based on its attributes, which can then be used as a starting point for negotiations between buyer and seller. The study shows the price prediction model in terms of three algorithms: Linear Regression, Random Forest Regressor and LightGBM using one main method to evaluation: Mean absolute error. Processing time is also another method for assessment, but it is a sub evaluation method. Many studies have been carried out to enhance the performance of models, such as removing outlier values, PCA, and so on. Feature selection in collaboration with hyperparameters tuning is considered as one of the best solutions, and LightGBM is the leader of three algorithms that always performs at a good level, no matter that before or after parameter tuning. Moreover, the processing time of LightGBM is always around one minutes and while tuning parameters it takes less time than Random Forest although the number of combinations is larger than Random Forest. The result found is consistent with previous findings of the key reference (Chada, 2019), as well as other research in the literature (Ge et al., 2020; Ke et al., 2017; Ponsam et al., 2021).

This can be explained that because LightGBM aggregates multiple prediction models from a more developed set of features with properly-tuned parameters to extract final result, so it can reduce the noise as well as perform faster and more efficiently compared to other algorithms, even though it has the tendency to fall into overfitting situation. The next consideration is to divide the dataset into small pieces according to the features or characteristics of each feature to achieve the better solution in each cluster. Here, clustering is not advised since it does not help to distinguish characteristic of each cluster, thereby manually dividing the dataset based on specific features is appropriate. Some features do not have significant impact on the models while the others are quite important that have a great influence on the result. Therefore, it should examine deeply before eliminating any features to exclude poor impacts to the model. In the future, it can be suggested that that future research can focus on testing and comparing performance of the boosting techniques, such as Gradient Boosting Machine, Gradient Boosting Machine, CatBoost, and ExtremeGBM (XGBM).

References

A Ramezan, C., A Warner, T., and E Maxwell, A. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, *11*(2), 185, 2019.

Akter, S., and Wamba, S. F. Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, *26*(2), 173-194. 2016.

Bro, R., and Smilde, A. K. Principal component analysis. *Analytical methods*, *6*(9), 2812-2831, 2014.

Chada, A. R. Strategic Pricing of Used Products for e-Commerce Sites. 2019.

Chandrashekhara, K., Thungamani, M., Gireesh Babu, C., and Manjunath, T. Smartphone price prediction in retail industry using machine learning techniques. In *Emerging Research in Electronics, Computer Science and Technology* (pp. 363-373), 2019.

Fathalla, A., Salah, A., Li, K., Li, K., and Francesco, P. Deep end-to-end learning for price prediction of second-hand items. *Knowledge and Information Systems*, *62*(12), 4541-4568. 2020.

Ge, D., Gu, J., Chang, S., and Cai, J. Credit card fraud detection using lightgbm model. *2020 International Conference on E-Commerce and Internet Technology (ECIT)*, 2020.

Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media. 2019.

Grgić, V., Mušić, D., and Babović, E. Model for predicting heart failure using Random Forest and Logistic Regression algorithms. *IOP Conference Series: Materials Science and Engineering*, 2021.

Gupta, R., and Pathak, C. A machine learning framework for predicting purchase by online customers based on dynamic pricing. *Procedia Computer Science*, *36*, 599-605. 2014.

Hurtado, P. A., Dorneles, C., and Frazzon, E. Big Data application for E-commerce's Logistics: A research assessment and conceptual model. *IFAC-PapersOnLine*, *52*(13), 838-843. 2019.

Jia, L., Zhao, Q., and Tong, L. Retail pricing for stochastic demand with unknown parameters: An online machine learning approach. *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154. 2017.

Kumar, V., and Minz, S. Feature selection: a literature review. *SmartCR*, *4*(3), 211-229. 2014.

Lenhart, T., Eckhardt, K., Fohrer, N., and Frede, H.-G. Comparison of two different approaches of sensitivity analysis. *Physics and Chemistry of the Earth, Parts A/B/C*, *27*(9-10), 645-654. 2002.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 1-45. 2017.

Minastireanu, E.-A., and Mesnita, G. Light gbm machine learning algorithm to online click fraud detection. *J. Inform. Assur. Cybersecur*, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press. 2018.

Narayana, C. V., Likhitha, C. L., Bademiya, S., and Kusumanjali, K. Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business. *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021.

Paper, D., and Paper, D. Scikit-Learn Classifier Tuning from Simple Training Sets. *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*, 137-163. 2020.

Ponsam, J. G., Gracia, S. J. B., Geetha, G., Karpaselvi, S., and Nimala, K. Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 2021.

Pundir, A. K., Ganapathy, L., Maheshwari, P., and Kumar, M. N. Machine learning for revenue forecasting: A case study in retail business. *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020.

Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., and Ghifari, A. Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. *DEStech Transactions on Computer Science and Engineering*, *10*. 2017.

Refaeilzadeh, P., Tang, L., and Liu, H. Cross-validation. *Encyclopedia of database systems*, *5*, 532-538. 2009.

Schaffer, C. Selecting a classification method by cross-validation. *Machine Learning*, *13*(1), 135-143. 1993.

Tortorelli, D. A., and Michaleris, P. Design sensitivity analysis: overview and review. *Inverse problems in Engineering*, *1*(1), 71-105. 1994.

Tziridis, K., Kalampokas, T., Papakostas, G. A., and Diamantaras, K. I. Airfare prices prediction using machine learning techniques. *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017.

Walfish, S. A review of statistical outlier methods. *Pharmaceutical technology*, *30*(11), 82. 2006.

Willmott, C. J., and Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82. 2005.

Yu, L., Zhou, R., Chen, R., and Lai, K. K. Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade*, *58*(2), 472-482. 2022.

Biographies

Nguyen Tran Quang Linh received his bachelor's degree in Logistics and Supply chain management at School of Industrial Engineering and Management, International University - Vietnam National University of HCMC.

MSc. Ngo Thi Thao Uyen is a lecturer in School of Industrial Engineering and Management, International University - Vietnam National University of HCMC. She received her master's degree from the Molde University College, Norway. Her research interests are international logistics and transportation, Data analytics in Logistics, Logistics games, Supply chain design, Supply chain coordination.

Mr. Duong Vo Nhi Anh is a lecturer in School of Industrial Engineering and Management, International University - Vietnam National University of HCMC. His research interests are Production Planning, Logistics and Supply chain management, Material planning.

Nguyen Thi Phuong Hoang is a undergraduate student majoring in Logistics and Supply chain management at School of Industrial Engineering and Management, International University - Vietnam National University of HCMC.