BOĞAZİÇİ UNIVERSITY

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

EXECUTIVE SUMMARY OF THE PRICE RECOMMENDATION SYSTEM FOR ONLINE RETAIL SELLERS

Ceren Öyke
Görkem Çoklar
Hakan Utku Özdemir
Zeynep Günce Gündüz

July, 2021

To begin our initiative, we collaborated with one of Turkey's leading second-hand marketplace, with the goal of providing a tailored shopping experience enhanced by modern technology. Most of the online marketplaces allow their sellers to price their products without any restriction. This is a more serious problem for second-hand shopping platforms since second-hand platforms have a C2C business model. Their sellers may list almost any product and also, most of these products are unique, or somewhat different in terms of condition, quality etc. Therefore, controls on pricing in second-hand platforms are more challenging. This study aimed to enhance the platform's health by providing more accurate price recommendations to sellers based on sold items using regression models, ensuring that both consumers and sellers are pleased with product pricing.

Projects steps include;

- Understanding established projects for price recommendation
- Conducting Literature Review
- Collecting data
- Preprocessing data
- Applying regression models
- Selecting the optimum model for suggestion by using evaluation metrics

In order to understand the established projects for price recommendation, an interview was conducted with a Data Scientist of the collaborated company. According to this interview, the price recommendation they made in the product submission page uses a markup percentage based on the original price of the product. However, this markup is predetermined so, it neither fully reflects the platform’s selling behavior nor is the optimum selling price for the platform's health.

To have an idea about the previous studies, a literature review is conducted. According to findings, a Random Forest model was used to predict house prices. Moreover, there were two other studies about price recommendation systems for second-hand platforms. One of them used image processing and conducted a vision-based regression model, while the other used various regression models. The latter study was based on a competition from the Kaggle platform.

With the help of their Data Analytics team, data for this project was obtained from a second-hand e-commerce site in Turkey. The information pertains to items that were sold between February 16th and June 16th, 2021. There are a total of 2.556.419 rows and 18 columns, indicating distinct sold goods. 5 columns are excluded from this project since four of them were ID’s of sellers, products, categories, and brands, and the other one was product status which is ‘SOLD’ in all the rows in the dataset. The data was mostly clear, understandable, and does not include missing, noisy or inconsistent data. Only 83.092 duplicate rows are deleted during the data preprocessing. Moreover, outliers should have been removed to increase the accuracy of the model. Therefore, the IQR method was used to eliminate outliers from our data. Since this study is planned to do by grouping based on the

brand title and the third-level category which is the deepest category level, the outlier elimination was done in these small data groups instead of doing it in the whole data. In this way, making selling price estimates for these groups, which is not affected by outliers is aimed.

From the correlation analysis, a weak negative correlation between “star” and “sold price” (-0.039) is observed. Thus, the variable of star that indicates the star ranking of a seller is not included in the model. The other continuous variables, price and original_price have quite a stronger correlation with the sold price. Since the price variable corresponds to the listing price of the product (price the users enter), it is not included in the model. On the other hand, new variables are generated to benefit from the categorical variables. Mean values of sold and original price for each group of brand and third level title are calculated separately. According to correlation analysis, sold and original price means by brand and third level titles has strong positive correlation between the sold price variable, so they are included in the model. Moreover, another variable called the discount rate is created using the formula (Original Price - Sold Price) / Original Price and again the average value of the discount rate is calculated according to brand title and third level title.

During the exploratory data analysis, it is discovered that some continuous variables in the dataset are right skewed. Therefore, to convert the non-normal distribution to a normal distribution, two normalization techniques are used: log and box cox transformation. The values normalized and adjusted to a common scale includes: sold price mean by third level title/brand title, original price mean by third level title/brand title. Log transformation performed better than box cox transformation with higher R^2 and lower RMSE. However, in the best performing model, normalization is not applied.

And finally, the models are created based on 3 regression models: Linear (OLS), Ridge and Lasso. In the Ordinary Least Squares regression, the model minimizes the squared errors. In the Ridge and Lasso models, however, the model minimizes different functions with a shrinkage penalty. This way, Ridge and Lasso models perform regularization and an alternative to subset selection. From the theoretical knowledge, it was expected ridge regressions to have higher prediction accuracy then the Lasso regressions. However, there were different possible combinations to create well performing models. Therefore, some functions are created using Python to run models easily and to be more flexible while generating these models. This way, normalizing and cleaning data, generating variables, creating models and calculating the evaluation metrics were possible just by changing and running a few lines of code. Therefore, this study is carried out using Python software language and its Pandas, Numpy, Seaborn, Sklearn, Matplotlib, Statsmodels, Scipy and Wordcloud libraries for various data science tasks on the Jupyter Notebook application which is an open source development environment.

In the best performing model with the highest R^2 and the lowest RMSE value, original and sold price outliers in both third level and brand title groups were determined and eliminated according to the IQR approach. In this model, 7 continuous variables and 9 dummy variables

were inserted into the model. While using the mean of the original price and sold price in different groups, discount rate, etc. as continuous variables; the product's status, shipment term, brand type in dummy format are used as categorical variables. The values of 0.657 for R^2, 0.657 for adjusted R^2, 633.367 for MSE, 25.167 for RMSE, and 0.439 for RMSLE are obtained. RMSE score of 25 for the Model 16 means that the model is predicting the sold price with +/- 25TL error rate. On the other hand, R^2 of 66% means the model explains 66% of the data.

To demonstrate a possible application of the model, a user interface is created. Users can simply enter the category, brand, brand type, shipment term, condition and original price of the product they want to sell and they can click the enter button. A recommended price will be generated by putting these inputs into the selected model.

This study will help sellers set reasonable and effective prices for their second-hand products without doing any market research. The customer perceives pricing to be the most important element in deciding whether or not to purchase used goods. Therefore, sales rates and profit will increase on the e-commerce platform thanks to the right price suggestion. With the implementation of our model into a real platform, the platform's health is expected to be improved due to more accurate price recommendations to sellers based on sold items, ensuring that both consumers and sellers are pleased with product pricing.

In further studies, gathering more data can train the model developed in this study better, thus can lead to more accurate predictions. Moreover, there may be additional variables that can be used as predictors depending on the platform using the model. On the other hand, Elastic Net regression which includes a combination of regularization techniques of Ridge and Lasso can be used to improve accuracy.