

CZ1115 Mini Project

Olist E-Commerce and Marketing
Datasets



Gideon Patrick Manik (U2021002K) , Rizwan Nusrath Fathima (U2022273F), Samiksha Sankar (U2021021D)
Lab Group FS7, Team 5

E-Commerce in Brazil

Earned USD27.6B in 2018

Projected to grow to USD38.5B by 2022

48M consumers in 2016

1/4 of Brazilians have purchased online at
least once



INTRODUCTION



How can
companies
increase their
revenue via e-
commerce?



How different variables affect the *review score* in different product type categories?

Why Review Score?

Research conducted by Profitero that customers are more willing to buy goods if sellers have high review scores

What are the variables?

Actual Delivery Time, Difference between Actual & Estimated Wait Time, Freight Value, Payment Value, Payment Installments

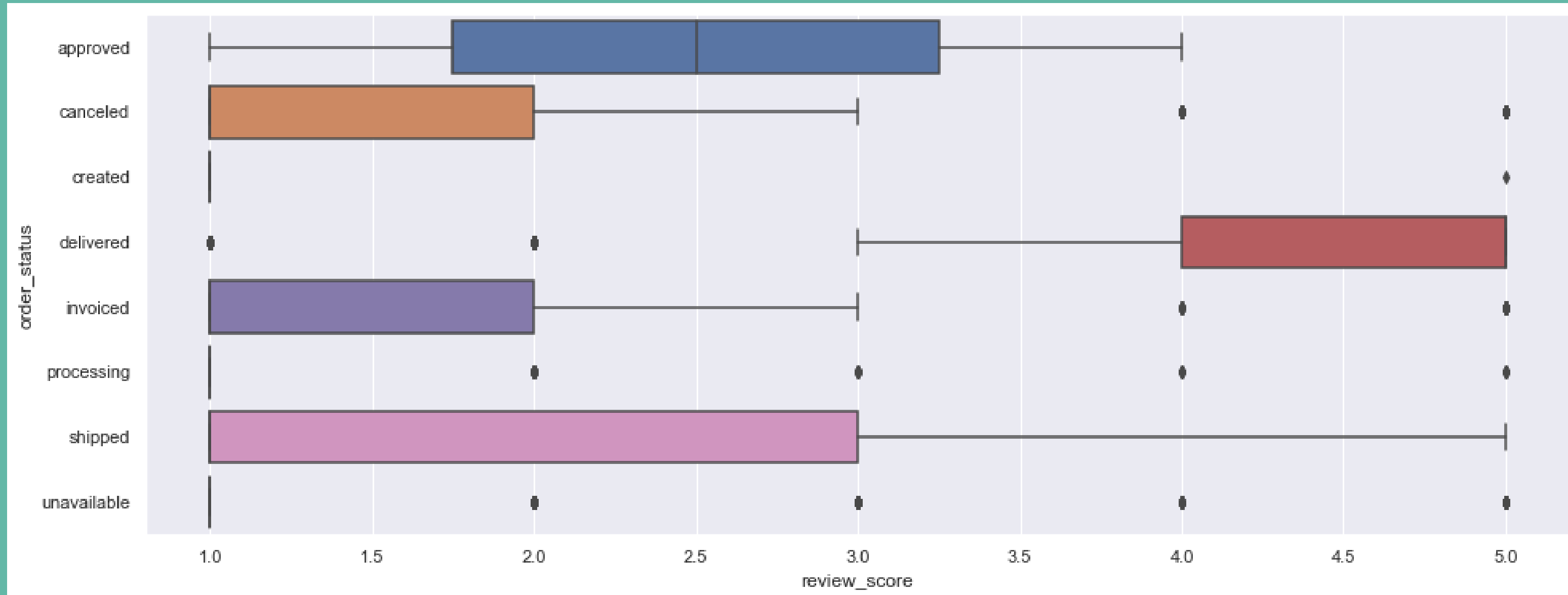




Exploratory Data Analysis

- Order Status
- Payment Value
- Payment Installments

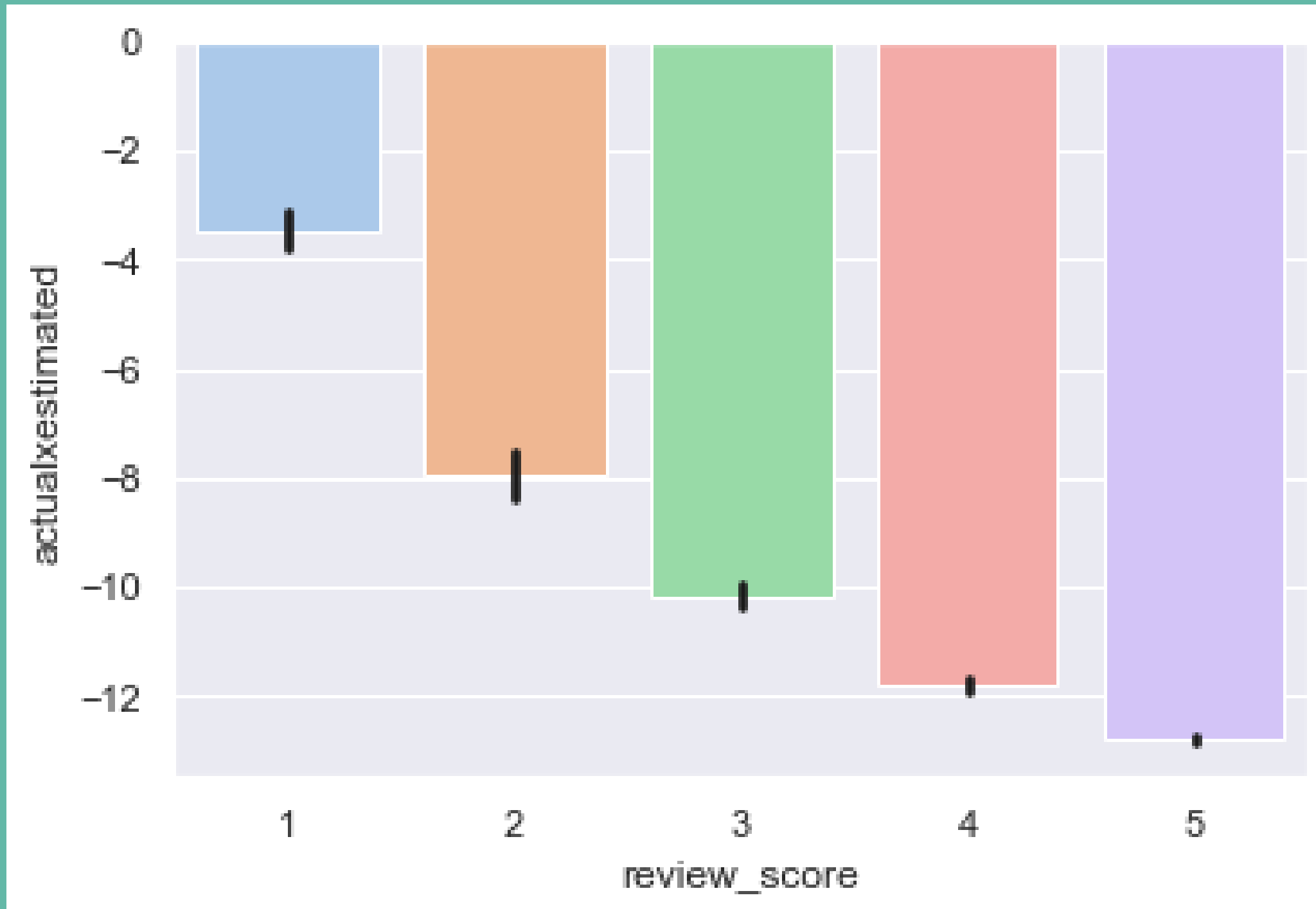
Order Status



97% of reviews are given to orders with delivered status, other statuses are too ambiguous to provide reliable review scores.

EXPLORATORY DATA ANALYSIS

Difference between Actual & Estimated Wait Time



Linear relationship, the more the delivery arrives earlier than expected, the higher the review

Actual Delivery Time



Linear relationship, the longer the delivery time, the lower the review.

Payment Value



There appears to be a correlation between high payment value and lower review scores. It can be used to distinguish between review scores 1, 2 and 3,4,5.

Payment Installments



No obvious correlation but one can expect to see correlation if payment installments is separated by product type as different product types have different characteristics.


How different variables such as Actual Delivery Time, Difference between Actual & Estimated Wait Time, Freight Value, Payment Value, Payment Installments affect the review score in each of the different product type categories, Houseware, Auto, Furniture Decor, Computer Accessories, Health & Beauty, Sports Leisure?

PROBLEM DEFINITION





Merging the datasets



Filtering the review scores based on order status to "delivered"



Reclassifying review scores



Splitting the dataset according to product categories



Remove duplicates

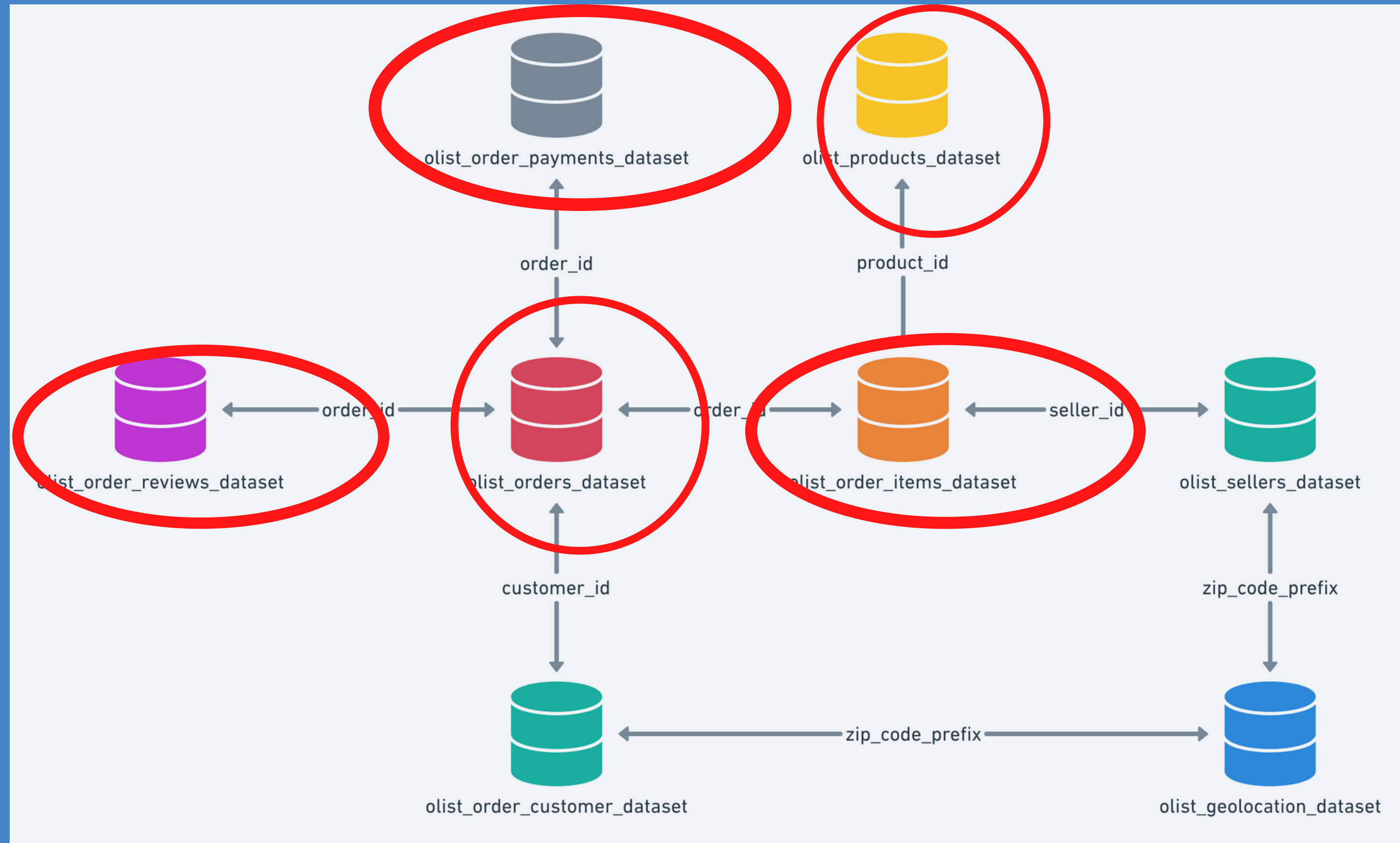


Remove null values



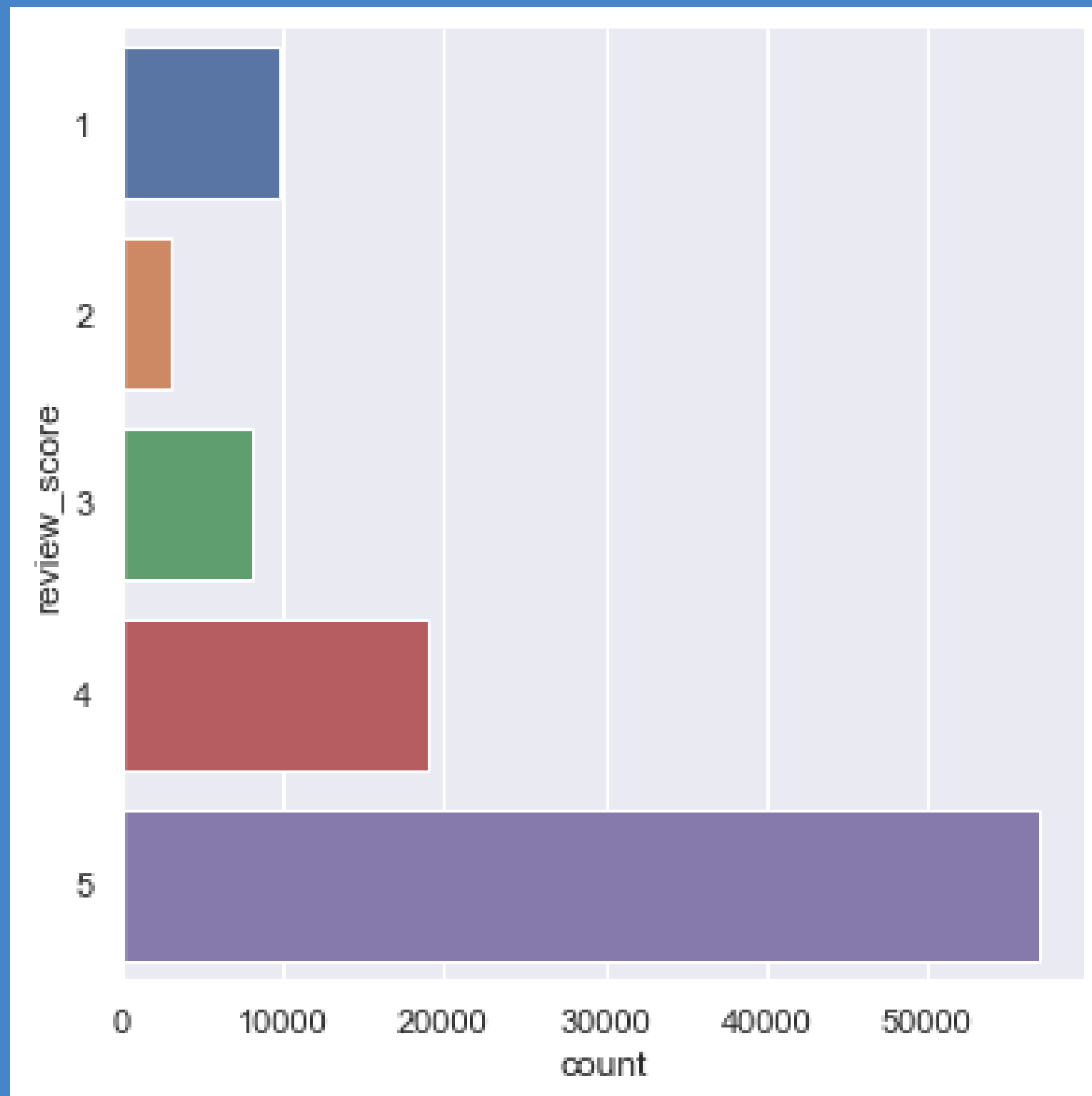
Balance review scores

Merging the datasets

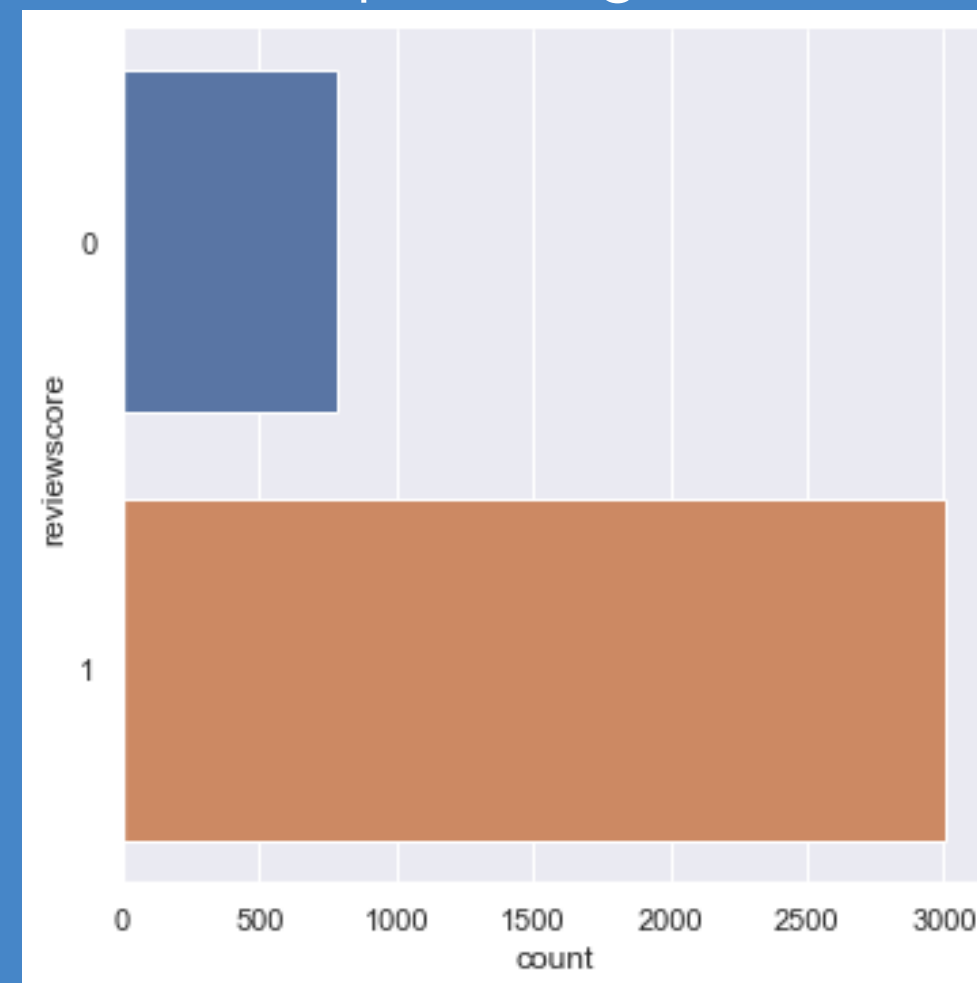


DATA EXTRACTION, CURATION, PREP & CLEANING

Reclassifying review scores



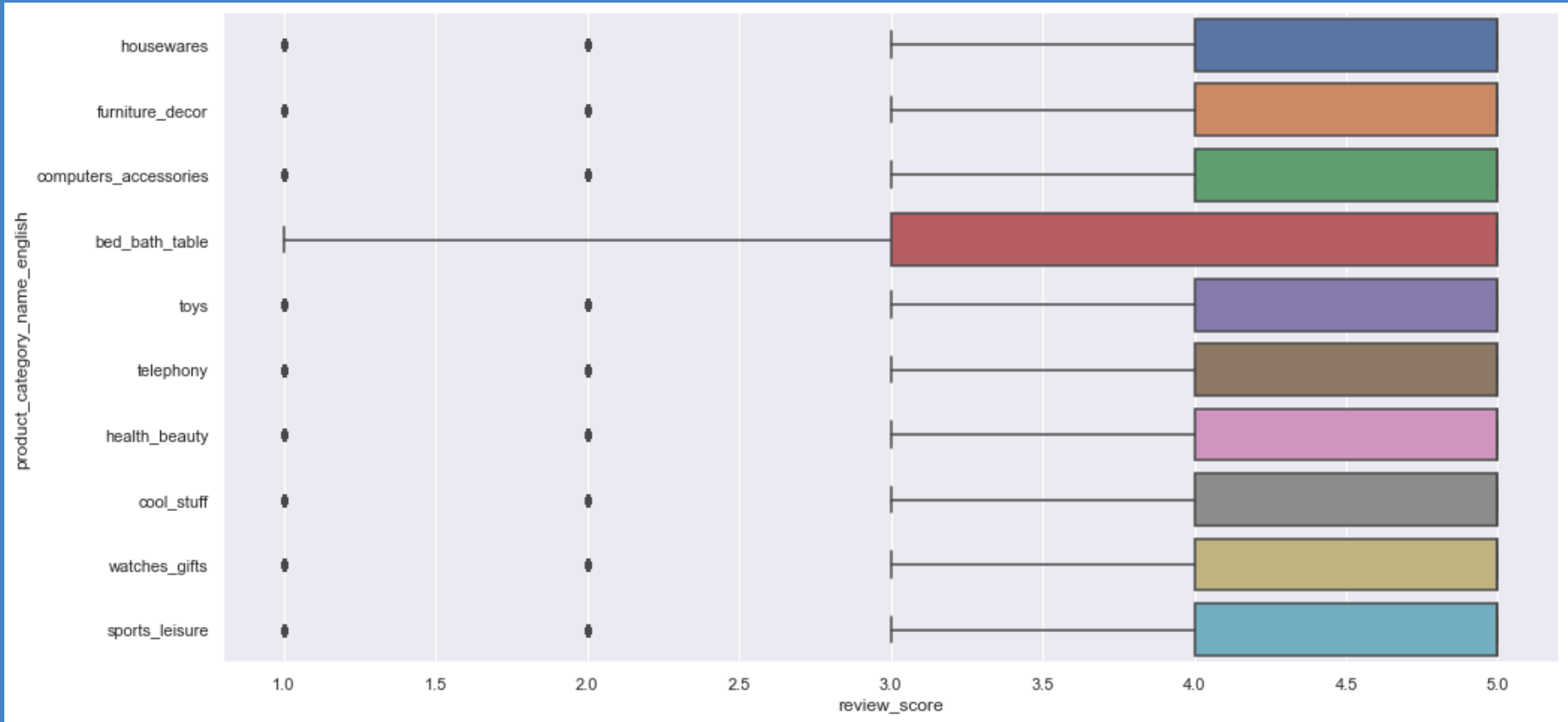
- Review score is an imbalanced categorical data type.
- Review score 1 to 3 classified as low (0). Review score 4 and 5 classified as high (1).
- Reclassification helps to reduce class imbalance without upscaling or downscaling



Splitting the dataset according to product categories

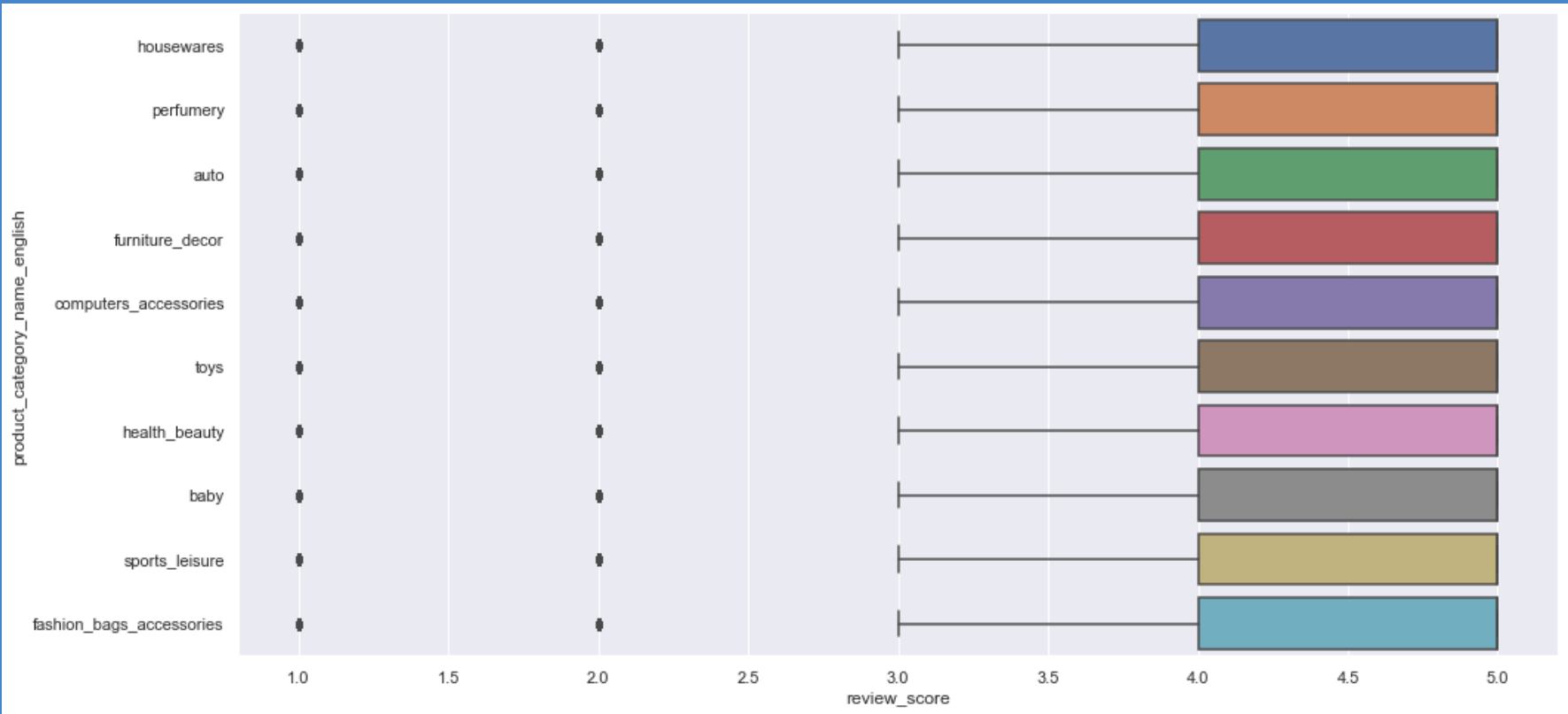
- Houseware, Auto, Furniture Decor, Computer Accessories, Health & Beauty, Sports Leisure
- These 6 categories were chosen as they were in the top 10 product categories for all 3 years.
- Review scores are split by product categories as each product type would have different characteristics that would cause variables like payment installments to be different.
- Data per category is more than enough to train the models.

What are the product categories?

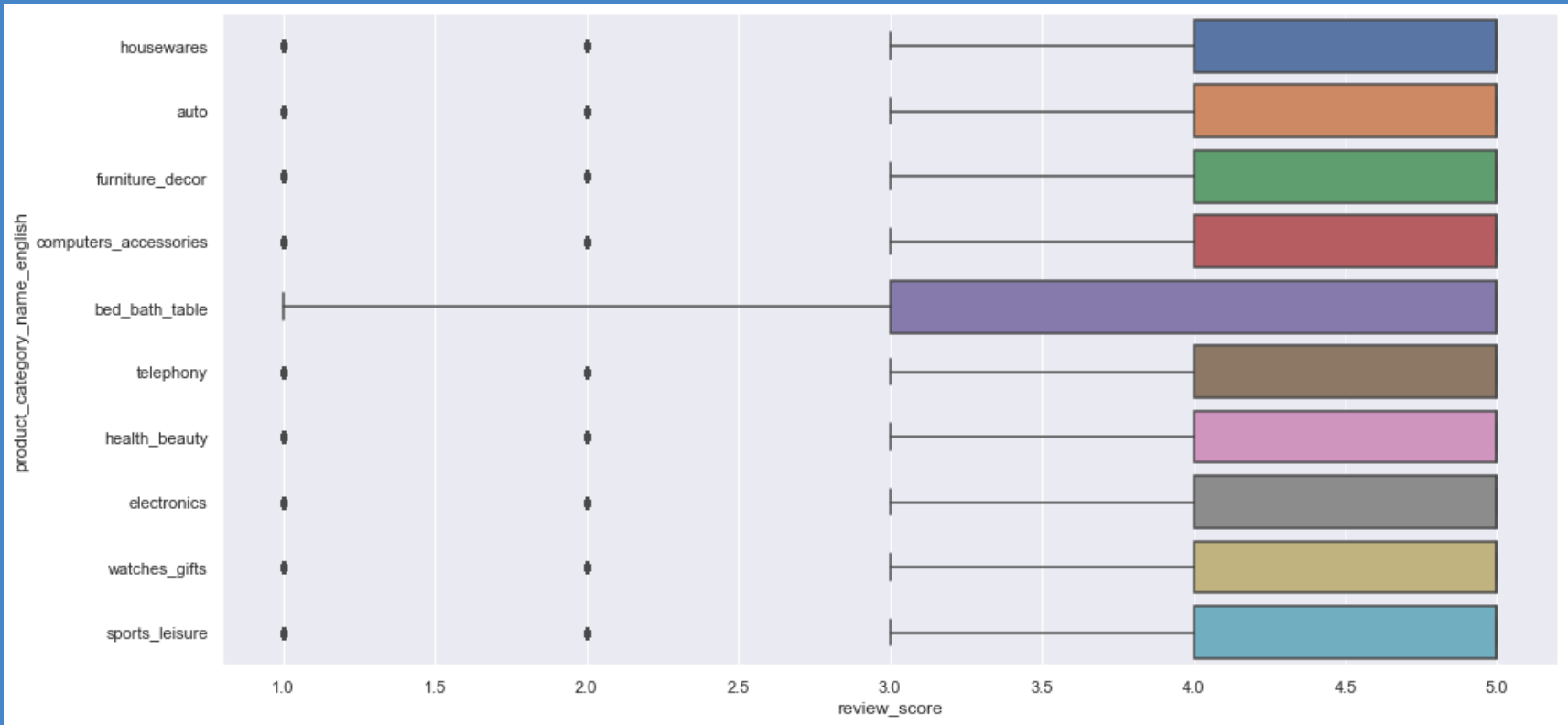


Top 10 Product Categories for 2017

Houseware, Auto, Furniture Decor,
Computer Accessories, Health & Beauty,
Sports Leisure



Top 10 Product Categories for 2016



Top 10 Product Categories for 2018

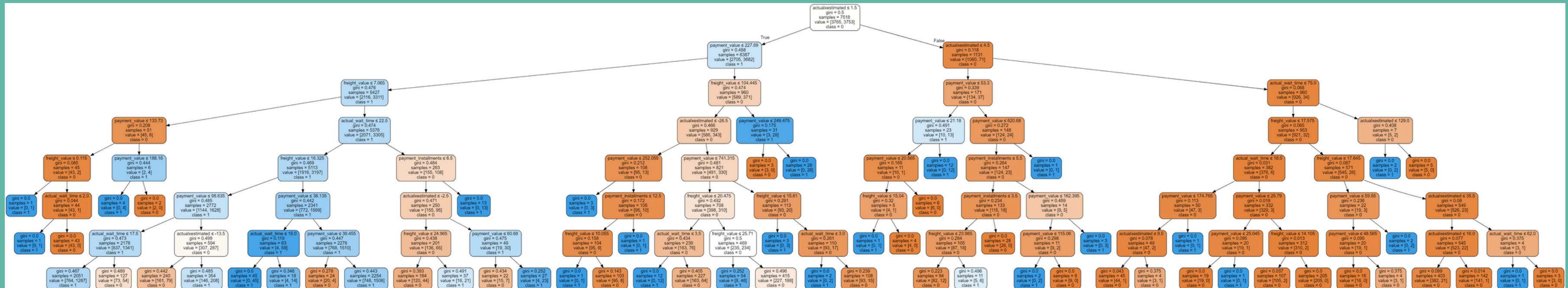


Machine Learning

- Decision Trees
- Random Forest
- GridSearch

Decision Tree

- A decision tree is used to classify a categorical response variable (review score) using numerical predictors.
- The classification accuracy was above 65% for most product types.
- False positive rate was below 50% for most product types.



Focusing on one type: HouseWare

Goodness of Fit of Model: Train Dataset

Classification Accuracy : 0.7078

True Negative Rate : 0.5498

True Positive Rate : 0.8662

False Negative Rate : 0.1338

False Positive Rate : 0.4502

Goodness of Fit of Model: Test Dataset

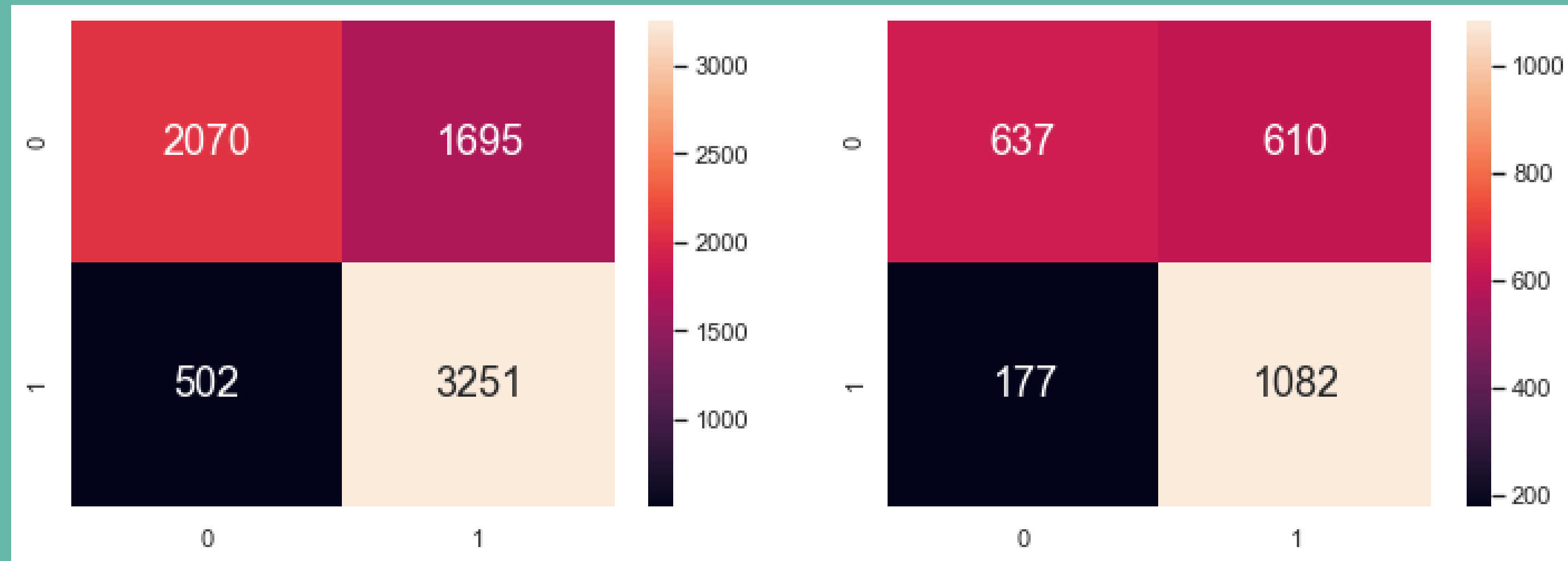
Classification Accuracy : 0.6860

True Negative Rate : 0.5108

True Positive Rate : 0.8594

False Negative Rate : 0.1406

False Positive Rate : 0.4892



Random Forest

- The classification accuracy increases to almost 80% for most product types.
- The false-positive rates drop below 30%.

Goodness of Fit of Model: Train Dataset

Classification Accuracy : 0.8176

True Negative Rate : 0.6781

True Positive Rate : 0.9574

False Negative Rate : 0.04260

False Positive Rate : 0.3219

Goodness of Fit of Model: Test Dataset

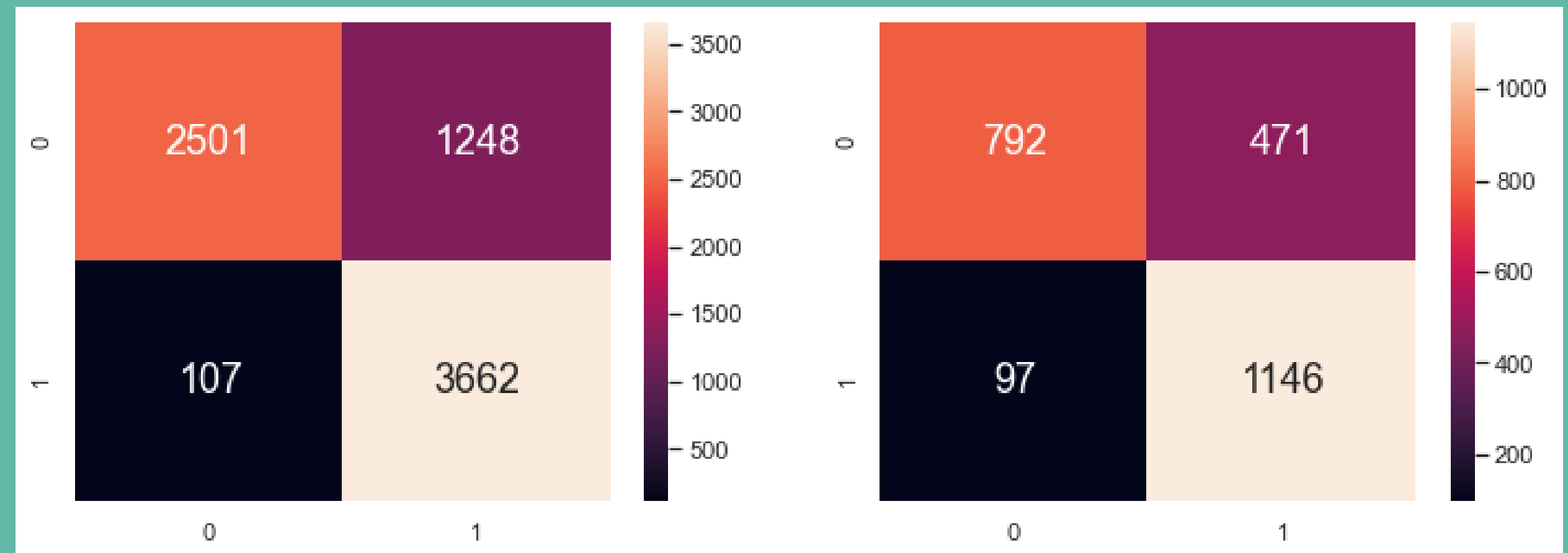
Classification Accuracy : 0.7905

True Negative Rate : 0.6600

True Positive Rate : 0.9204

False Negative Rate : 0.0796

False Positive Rate : 0.3400



Cross-Validation GridSearch Results

- GridSearch was run over two different ranges `np.arange(9, 14)` and `np.arange(2, 11)`
- This was to find the best hyperparameters.
- Let's look at the confusion matrix HouseWare at a `max_depth` of 10 and `n_estimators` 400 (values were obtained after GridSearch)

Product Category	Furniture		Health and Beauty		Auto		Computer Accessories		Sports and Leisure		HouseWare	
max_depth	10	13	10	13	10	13	10	13	10	13	10	13
n_estimator	200	900	800	800	500	300	500	700	800	900	400	600
best_score (to 3 decimal places)	0.811	0.893	0.807	0.929	0.822	0.904	0.796	0.863	0.805	0.869	0.786	0.876

After Grid Search Results

Goodness of Fit of Model: Train Dataset

Classification Accuracy : 0.9953

True Negative Rate : 1.000

True Positive Rate : 0.9907

False Negative Rate : 0.0093

False Positive Rate : 0.00

Goodness of Fit of Model: Test Dataset

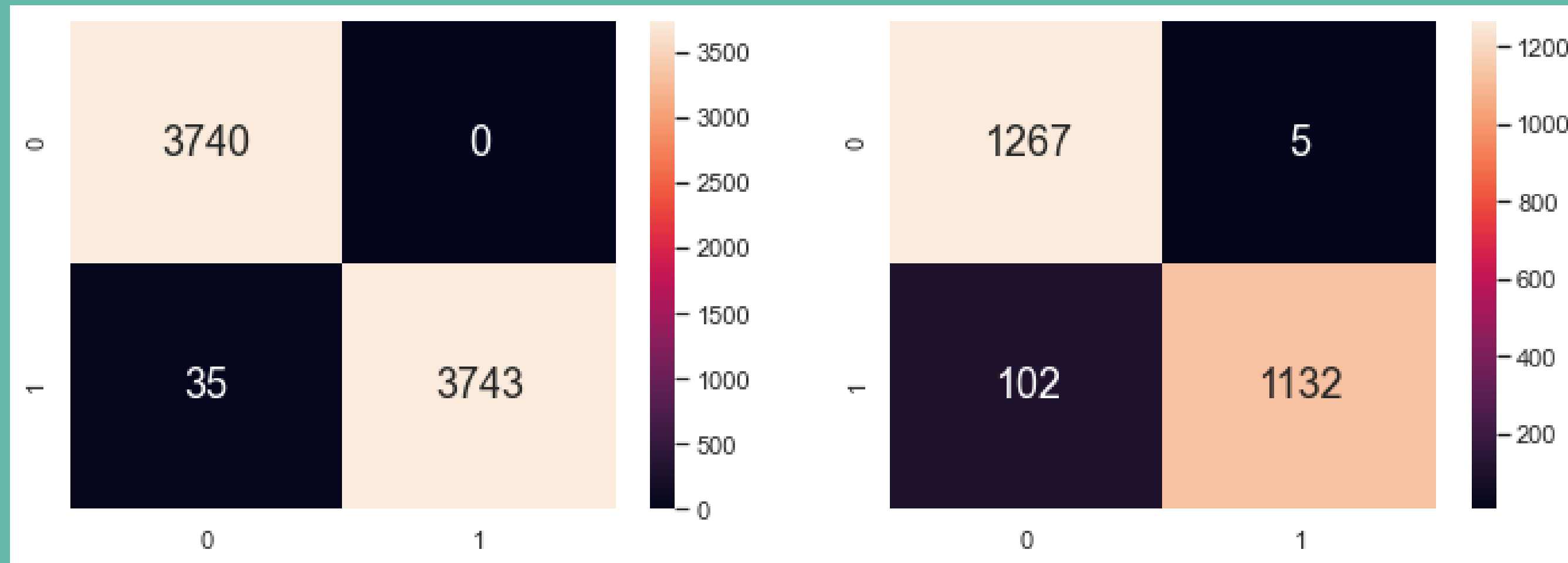
Classification Accuracy : 0.9573

True Negative Rate : 0.9961

True Positive Rate : 0.9173

False Negative Rate : 0.0826

False Positive Rate : 0.0039



Comparing Results From the CV Grid searches

Product Category	Furniture			Health and Beauty			Auto		
	Decision	RF	After Grid Search	Decision	RF	After Grid Search	Decision	RF	After Grid Search
Classification Accuracy	0.6797	0.8033	0.923	0.7057	0.8106	0.959	0.7192	0.8156	0.954
True Positive	0.6576	0.8758	0.857	0.8106	0.9352	0.925	0.5545	0.9500	0.909
True Negative	0.7017	0.7307	0.992	0.6025	0.6811	0.992	0.8919	0.6828	0.999
False Positive	0.2983	0.2693	0.008	0.3975	0.3189	0.008	0.1081	0.3172	0.001
False Negative	0.3424	0.1242	0.143	0.1894	0.0648	0.075	0.4455	0.0500	0.091
Product Category	Computer Accessories			Sports and Leisure			HouseWare		
	Decision	RF	After Grid Search	Decision	RF	After Grid Search	Decision	RF	After Grid Search
Classification Accuracy	0.6989	0.8032	0.928	0.7312	0.7978	0.955	0.6860	0.7905	0.957
True Positive	0.8306	0.8916	0.870	0.8941	0.9482	0.916	0.8594	0.9204	0.917
True Negative	0.5731	0.7145	0.987	0.5525	0.6508	0.995	0.5108	0.6600	0.996
False Positive	0.4269	0.2855	0.013	0.4475	0.3492	0.005	0.4892	0.3400	0.004
False Negative	0.1694	0.1084	0.130	0.1059	0.0518	0.084	0.1406	0.0796	0.083

What we learned

Using a different Machine Learning Function

- Random Forest
- Cross Validation GridSearch

Outcome of Project

Using our ML function, Sellers can:

- Predict their review scores based on their business model
- Modify variables & predict what changes can be made to improve their review score

CONCLUSION

In conclusion, these are the data driven insights

- Lower actual wait time and actual - estimated wait time
liaise with delivery team to deliver faster
- Lower Freight Value
order in bulk or find companies that allow lower freight value
- Lower Payment value
Create deals with credit card companies or banks to enable the lowest payment value
- Lower payment instalment values
Create better instalment plans

CONCLUSION

Thank You!

Gideon Patrick Manik (U2021002K) :
Script

Rizwan Nusrath Fathima (U2022273F):
Slides

Sankar Samiksha (U2021021D):
Coding



CONCLUSION

Reference List

1. Olist, “Brazilian E-Commerce Public Dataset by Olist,” Kaggle, 29-Nov-2018. [Online]. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_orders_dataset.csv. [Accessed: 23-Apr-2021].
2. L. Breiman and A. Cutler, “Random Forests Leo Breiman and Adele Cutler,” Random forests – classification description. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. [Accessed: 23-Apr-2021].
3. K. Anderson, D. Reifenberger, and T. O' Neil, “Assessing the Impact of Ratings and Reviews on eCommerce Performance,” <https://www.profitero.com/>. [Online]. Available: <http://insights.profitero.com/rs/476-BCC-343/images/Assessing%20the%20Impact%20of%20Ratings%20and%20Reviews%20on%20eCommerce%20Performance.pdf>. [Accessed: 23-Apr-2021].
4. “eCommerce is Set to Increase 39 Percent by 2022 in Brazil, Reaching Nearly R\$150bn – Press Releases: FIS,” FIS Global. [Online]. Available: <https://www.fisglobal.com/en/about-us/media-room/press-release/2018/ecommerce-is-set-to-increase-39-percent-by-2022-in-brazil-reaching-nearly-r150bn>. [Accessed: 23-Apr-2021].
5. B. L. F. @biancamvickers, “Brazilian E-commerce Market 2016 Highlights,” PagBrasil, 17-Jul-2019. [Online]. Available: <https://www.pagbrasil.com/insights/brazilian-e-commerce-market-2016-highlights/>. [Accessed: 23-Apr-2021].

CONCLUSION

Reference List

6. H. M. –, By, –, Hussain MujtabaHussain is a computer science engineer who specializes in the field of Machine Learning.He is a freelance programmer and fancies trekking, H. Mujtaba, Hussain is a computer science engineer who specializes in the field of Machine Learning.He is a freelance programmer and fancies trekking, and P. enter your name here, “What is Cross Validation in Machine learning? Types of Cross Validation,” GreatLearning Blog: Free Resources what Matters to shape your Career!, 24-Sep-2020. [Online]. Available: <https://www.mygreatlearning.com/blog/cross-validation/>. [Accessed: 21-Apr-2021].
7. abuabu 54777 silver badges1616 bronze badges, Mischa LisovyiMischa Lisovyi 2, and Vivek KumarVivek Kumar 28.8k66 gold badges7575 silver badges109109 bronze badges, “Interpreting sklearn's GridSearchCV best score,” Stack Overflow, 01-Feb-1967. [Online]. Available: <https://stackoverflow.com/questions/50232599/interpreting-sklearn-gridsearchcv-best-score>. [Accessed: 21-Apr-2021].
8. R. Joseph, “Grid Search for model tuning,” Medium, 29-Dec-2018. [Online]. Available: <https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e>. [Accessed: 21-Apr-2021].
9. Shahul ES Freelance Data Scientist | Kaggle Master Data science professional with a strong end to end data science/machine learning and deep learning (NLP) skills. Experienced working in a Data Science/ML Engineer role in multiple startups. K, S. ES, Freelance Data Scientist | Kaggle Master Data science professional with a strong end to end data science/machine learning and deep learning (NLP) skills. Experienced working in a Data Science/ML Engineer role in multiple startups. Kaggle Kernels Master ra, and F. me on, “Hyperparameter Tuning in Python: a Complete Guide 2021,” neptune.ai, 19-Mar-2021. [Online]. Available: <https://neptune.ai/blog/hyperparameter-tuning-in-python-a-complete-guide-2020#:~:text=Hyperparameter%20tuning%20is%20the%20process,maximum%20performance%20out%20of%20models>. [Accessed: 21-Apr-2021].

CONCLUSION

Reference List

10. J. Brownlee, "Hyperparameter Optimization With Random Search and Grid Search," *Machine Learning Mastery*, 18-Sep-2020. [Online]. Available: <https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/>. [Accessed: 21-Apr-2021].
11. "sklearn.model_selection.GridSearchCV¶," *scikit-learn*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed: 21-Apr-2021].
12. M. Sharma, "Grid Search for Hyperparameter Tuning," *Medium*, 21-Mar-2020. [Online]. Available: <https://towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec>. [Accessed: 21-Apr-2021].
13. R. Meinert, "Optimizing Hyperparameters in Random Forest Classification," *Medium*, 07-Jun-2019. [Online]. Available: <https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6>. [Accessed: 21-Apr-2021].
14. "sklearn.ensemble.randomforestclassifier¶." [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. [Accessed: 21-Apr-2021].
15. Slides & All designs from Canva