Milestone 2

Draft of White Paper

Jocelyn Disla

DSC 680-T302 Applied Data Science

Amirfarrokh Iranitalab

Mar 31, 2024

**The Problem**

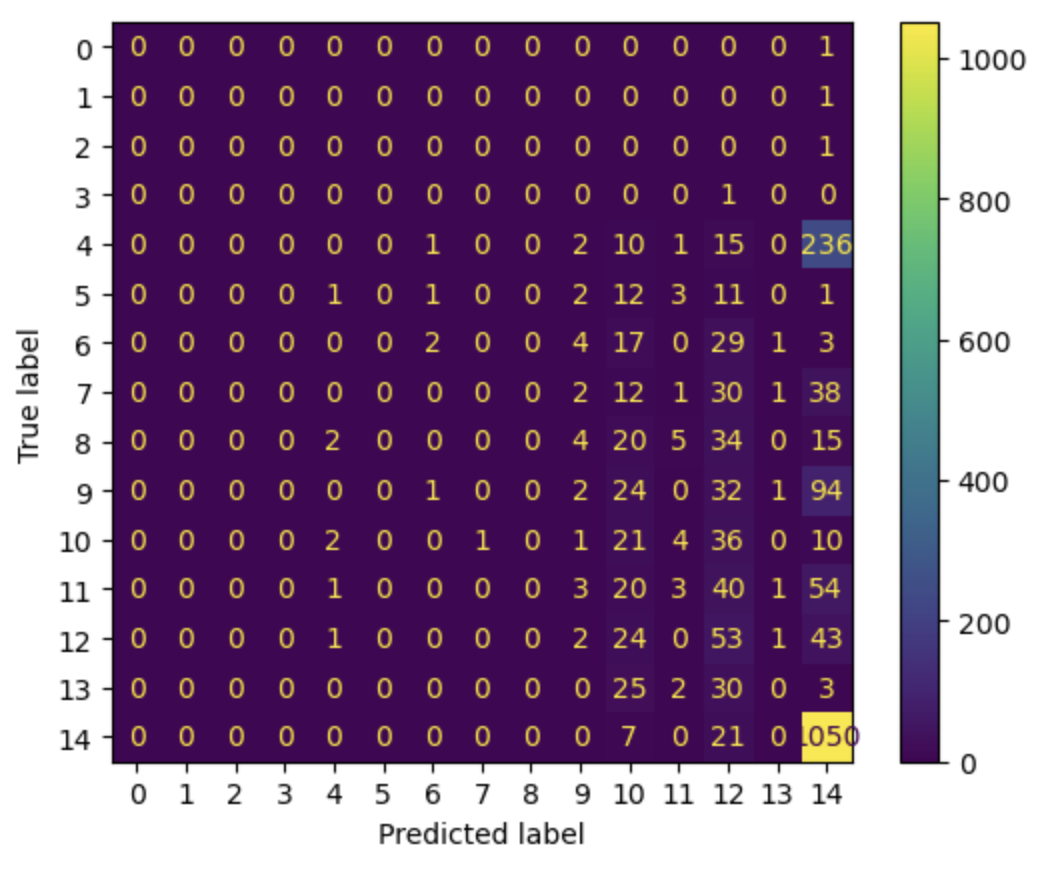
A predicitve system for a business selling any form of a product is very useful to gather information based on consumer feedback to understand which products will be bought and desired most. This system will be called structured advice, for its helpful suggestions to marketing teams and production for its ability to predict sales and overall product favorability.

Businesses typically have the issue of having so many clients or consumers that many come unnoticed, an automatic system will provide insights to be able to predict which types of products will perform the best according to customer’s desires. Every business also strives to continuously maximize their profits, by adding a predictive system there comes comfort to consumers to trust the company since they know them so well, building credibility. Many times another observable issue for companies is following bad predictive advice, this can potentially lead to loss of revenue and profits. With a predictive system, this issue is combatted by using customer feedback and machine learning to distinguish the products that will be the most attractive.

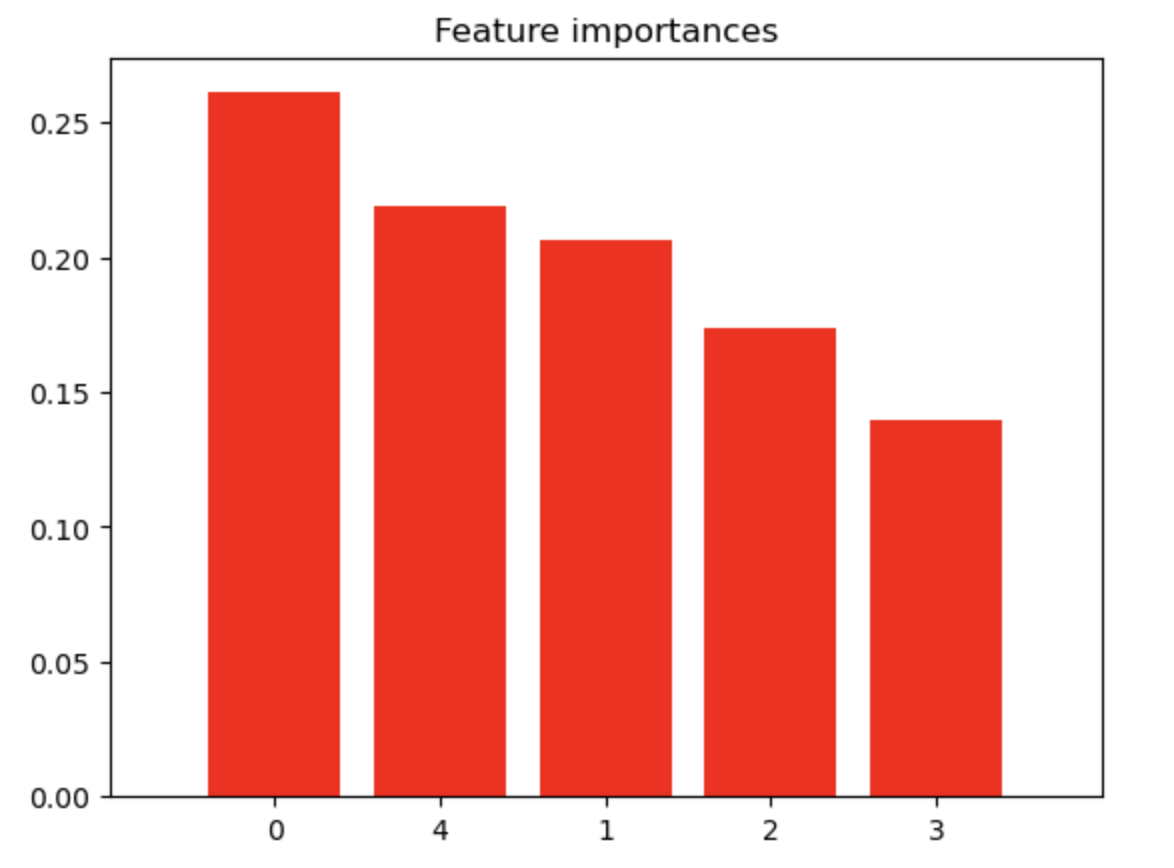
**Data, Method, and Results Explanation**

The dataset from Data.world was retrieved and used to create the predictive system, creating structured advice. This dataset is from Amazon fashion sales that describes the product, its rated review, manufacturer, price, and additional product information. This dataset would be a very large dataset around 10,000 rows with 17 features. In the process of cleaning the data, the features that were used for the building the model were the product name, category, rating, number of reviews, maker and price. The rating feature was cleaned up by removing excess wordage such as ‘our of 5 stars’ so that its numerical value is all that is used. The price is in euros and the symbol was removed similarly with a for loop in order to provide solely numerical values for the model to analyze the most impacting features. The string values in the product name, category, and maker were converted into distinct numerical values that are synced with their string value. The only feature that was left untouched was the number of reviews a product held. Once the nan values were removed from the dataset, the target variable was set as the review rating which was a float ranging from 0 to 5, 0 being the least liked and 5 being highly favorited. The features were the rest of the columns within the dataset, as seen in the appendix.

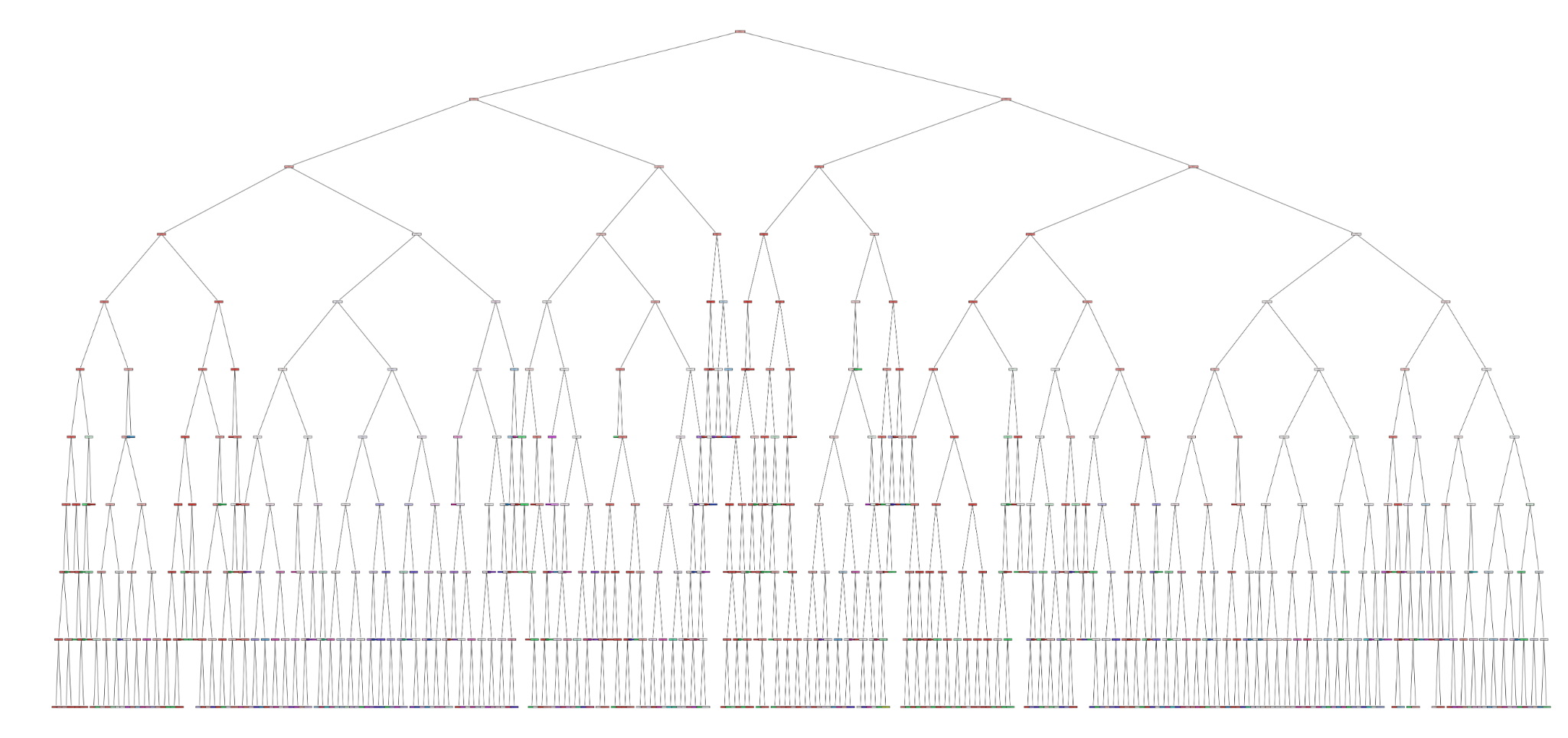
The data was split into a training and test set with the training set being 75% of the data and the remain 25% was used for the test set for the random forest model applied. The score of 0.538 for the training set using an out-of-bag estimate which signifies the number of correctly predicted rows. The oob score was not impressive and demonstrated a weak prediction of the results with a 3% rate above random guessing. The calculated accuracy of this model also provided a similar low score of 0.529, thus demonstrating the model did not learn anything useful with the features compared to the target variable. A confusion matrix was calculated for the model with a calculated hypothetical best model, the matrix demonstrated incongruence that supported the model’s lack of learning.



The feature importance displayed the number of reviews being the most impactful feature to the products rating and the product itself as the second most important feature. The more ratings a product has the more likely it is to have a higher rating, then the product itself is important to its own rating.



The least important feature was the category the product was in, meaning customers are not looking for a category first when deciding what product to buy. Lastly, the plotted trees emphasized the target variable’s non-binary values, since the ratings were floats there were many trees that made the plot much more complex, making it complex and difficult to understand.



**Conclusion and Assumptions**

The random forest model performed poorly with the amazon dataset, performing only slightly better than a random guess. Customers are focusing more on the product and reviewing the ratings others have left before deciding to buy. This suggests that customers are not looking through a category to find their desired items but rather looking up the name specifically then gathering information from previous buyers to decipher if buying the item would be a good idea or not. This information creates a theory that the more ratings an item has the more likely it will be continued to sought after. According to the Luke Burgis article regarding mimetic desires, seeing someone else have an item or enjoying an item makes it more desireable to the person observing. This supports the model’s feature importance calculation of number of ratings, since more people are buying this item there must be something amazing about it that has them buying and writing reviews on it. This model would suggest that businesses focus on retrieving serious public feedback on their products that can potentially be more alluring than traditional sales marketing tricks.

**Implementation, Future plans, and Recommendations**

Applying another model to this system would create a more efficient and trustworthy accuracy that companies would be able to make important profit decisions off of. In the future applying this model to products ratings would need more information regarding the review’s quality and overall message, such as the amount of wording placed in the review would have a bigger impact on its effect on passing viewers. Implementing this model for insightful findings can also bring new ideas to the table for marketing teams while opening new doors to organized data on reviews. Applying a tokenizer to the reviews so the machine can evaluate the messaged would be very beneficial moving forward.

**Ethical Implications and Challenges**

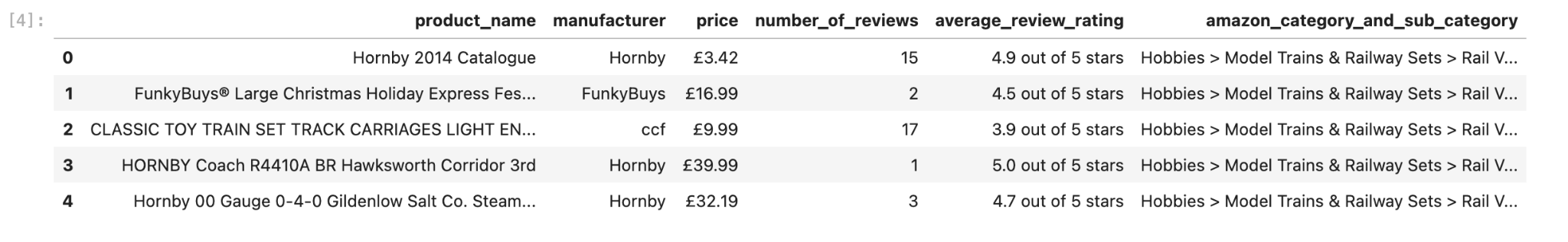
Some ethical implications stem from gathering information on highly reviewed items to influence a consumer to buy items that they haven’t consciously chosen or thought of. This intention would be to prey on consumers’ psychology and mental vulnerabilities which can create potential issues, ethically. Based on the data itself, Amazon products, reviews, descriptions, and other information are public and will not create any ethical issues.

Challenges in the future that may come around would be working with the tokenizer to evaluate the descriptions so that the outcome provides insights on recommending similar products. Another challenge would be to decide which feature to focus on so that the process is not as time-consuming while also still maintaining important features to maximize the insights from the models applied.

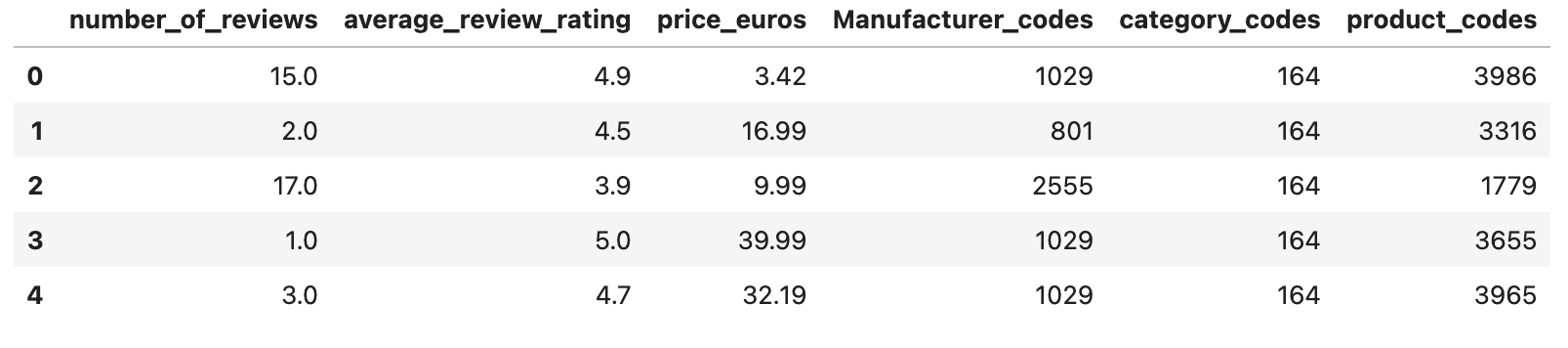
**Audience Questions**

Some potential questions that would arise would be; does applying this model waste more time than insights provided, since its accuracy is so low? How does additional features alter the model’s accuracy? If another model were to be added for a more accurate result, what model would you add? What would need to be done to increase the accuracy? Why did you decide to use a random forest model? If you had to pick a different target variable to help answer this problem, what would it be? Why did you decide to use codes for the columns with string values? Is there a better way to display your results? How confident are you in this model? Are there any other benefits or insights that can be retrieved using this model and data?

Appendix



Dataset transformed to equivalentnumerical values



References

*Fashion products on Amazon.com - dataset by promptcloud*. data.world. (2017,

September 18). https://data.world/promptcloud/fashion-products-on-amazon-com

*Know what your customers want before they do*. Harvard Business Review. (2014, October 28). https://hbr.org/2011/12/know-what-your-customers-want-before-they-do

*Mimetic desire*. Luke Burgis. (2023b, September 10).

https://lukeburgis.com/mimetic-desire/

R, S. E. (2024, January 3). *Understand random forest algorithms with examples (updated*

*2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=It%20can%20handle%20binary%2C%20continuous,robust%20model%20with%20some%20 limitations.&text=Random%20forest%20algorithm%20is%20an,to%20enhance%20a%20model’s%20performance.