Determining Fairness in the Gig Economy

Applied Data Science

CPSC 4300 – Spring 2023

Janet Taylor

Clemson University

Modeling, Cleaning, Writing

janett@clemson.edu

Evan Kessler

Clemson University

Sentiment Analysis, Labeling, Writing

ekessl@clemson.edu

**ABSTRACT**

Workers in the gig economy (i.e., Uber, GrubHub, TaskRabbit, etc.) depend on the reviews they get to continue working on the platform. Workers on those platforms are often affected by receiving negative reviews that describe elements that were not under their control. The project explores how the analysis of online reviews can detect when unfair reviews are left to workers.

# INTRODUCTION

Throughout the project, we posed the question “What factors contribute to unfair ratings in the gig economy and can we detect such metrics to limit their influence on worker’s performance?” To perform the analysis, we focused on two of the largest companies in the gig economy industry: Uber and Lyft. These two companies primarily focus on taxi services: riders are matched with drivers through the company’s app and driven to their requested location. Riders can leave reviews on their experience with their drivers, along with a rating that ranges between one and five stars. These reviews are extremely important for the driver, as anything other than a near-perfect rating causes the app to penalize the driver with fewer rides and may even result in their termination from the platform [1]. The objective of this project was to determine which reviews left to drivers could be determined as “unfair,” so that the company could disregard these reviews when calculating the driver’s overall rating. In total, 15,455 reviews were sourced from Uber and Lyft’s Trustpilot pages; 6,475 reviews were considered valid for training and testing. 3,771 reviews were for Uber and 2,704 reviews were for Lyft.

# EXPLORATORY DATA ANALYSIS

## Cleaning

One review corresponds to one row in the dataset. The relevant statistics for each review were the number of reviews previously left by the user, the review title, the review body, the date on which the ride occurred, the rating, whether the company requested the review be left, and the company that the review was for. The original dataset was collected using the Octoparse scraper, and thus required thorough cleaning. We attempted to filter out reviews that were targeted towards the company itself and/or were left by drivers, which filtered down the number of usable reviews to 6,475 unique observations with effective dates between January 2014 and February 2023. Only about two out of five negative reviews left on Trustpilot mentioned drivers. We also separated out numerical information buried in text (ex., the rating column was originally “Rated X out of 5 stars” rather than just the number we were interested in).

Table 1. The percentage of kept reviews for Uber, by rating

Graphical user interface

Description automatically generated

Table 2. The percentage of kept reviews for Lyft

Graphical user interface, text, application

Description automatically generated

## Initial Analysis

Before we could make more thorough insights about the fairness of the reviews, we first wanted to know the breakdown of the dataset. There were some immediate differences between Uber and Lyft datasets; namely, Lyft did not prompt users for reviews on Trustpilot. Our key predictors were review title, review body, and rating. Review title and review body were condensed into a “Sentiment” value to create a numerical predictor.

Chart, pie chart

Description automatically generated

Figure 1. The ratio of prompted to unprompted reviews for Uber

All reviews are heavily biased towards either one or five star, which is expected given the nature of ratings; generally, mediocre experiences do not incentivize people to leave the platform in order to leave a review. Figure 2 illustrates this. The number of ratings were normalized to the number of reviews present for the company, so the difference in quantity between the Uber and Lyft reviews was not reflected.

Chart, bar chart

Description automatically generated

Figure 2. The frequency of each rating in the dataset, by company

Figure 2 also illustrates that Lyft reviews tended to be more negative than Uber reviews. Lyft also had more reviews in the two-to-four-star range.

The ratings for each company were broken down into three categories. Reviews rated one or two stars were considered negative, reviews rated three stars were neutral, and four or five starred reviews were rated positive. Table 3 illustrates this.

Table 3. A summary of the breakdown of reviews by company

Table

Description automatically generated with medium confidence

## Sentiment Analysis

To get the sentiment of a review, the TextBlob library was used [2]. A sentiment of a review is the sum of the sentiment of all its words. Words were corrected for grammar and spelling in order to ensure uniform processing. Stop words and words used less than twenty times throughout the dataset were also excluded from the calculation.

Chart, bar chart

Description automatically generated

Figure 3. The top twenty words associated with the highest and lowest sentiment

We also analyzed sentiment trends for both companies. Uber’s sentiment seemed to be much worse in 2021 than 2022, and Lyft seems to strongly be associated with a negative sentiment after 2018. Our theory is that much of the sentiment might have changed with Uber after the COVID-19 pandemic.

Moreover, this change in sentiment month by month for both Lyft and Uber possibly indicate how the public feels about these certain companies at a given time. The business model has mostly stayed the same with drivers and customers. Drivers pick up the riders and drive them to their destination. There is variability in the prices and the apps, which possibly could have  
caused changes in how customers perceive these companies. Sentiment has changed for Lyft and Uber over time. People, on average, recently have had a negative experience when reviewing Lyft and a positive experience when reviewing Uber. Considering how people feel in general about these companies over time affects how one should consider fairness for the partners working in the gig economy. We should expect the sentiment to remain about the same month to month in the service that these companies provide. After all, there is a wide selection of drivers and many samples to be taken from these companies. Unless there has been a revolutionary shift on who is allowed to drive for Uber and Lyft, we hypothesize that much of the sentiment change is not about the drivers, but rather the business itself. Considering this shift, perhaps the companies can see what they have been doing to affect how people feel about the company rather than forcing drivers to absorb the reputation hit.

Chart, histogram

Description automatically generated

Figure 4. Average sentiment polarity by month for Uber Reviews. Sentiment polarity was determined using the TextBlob library. Each review was considered through its sentiment polarity and used in the calculation for the  
average sentiment polarity for the month.

Chart, histogram

Description automatically generated

Figure 5. Average sentiment polarity by month for Lyft Reviews. We used the same methodology to calculate the average sentiment polarity as we did for the Uber Reviews.

We also investigated if the sentiment was lower on specific days. We observed the average sentiment over numerous days of the year. Certain days, like May 10, 2021, (Mother’s Day) and the weekend before Halloween 2021 (October 27 and 28), were days with abnormally low sentiment. Reading into the reviews reveals that this was because there was a lack of drivers to handle the number of passengers looking for rides, which is outside of the control of individual drivers. To expand our model, one could determine if the date the review was placed was one with high traffic to the service and factor that into deciding if the model is unfair.

While interesting, including this date data into our models did not improve accuracy. We believe this is because more data and inputs are needed to use the time to extrapolate days with high stress on the system, not only just reviews, but perhaps some app usage statistics as well. Numerical statistics such as rides requested, rides denied in certain geographical areas, and the number of drivers operating in a region are crucial to understanding if a review is truly related to the driver’s performance or was unfairly left out of annoyance due to wait times from high demand. However, this baseline analysis does show that times that are more contested (weekend before Halloween, or Mother’s Day) can contribute to the sentiment of the customer leaving the review and may have an impact on fairness.

Chart, waterfall chart

Description automatically generated

Figure 6. Average sentiment polarity for Uber reviews for days with >5 reviews where the overall polarity was negative.

Chart, waterfall chart

Description automatically generated

Figure 7. Average sentiment polarity for Lyft reviews for days with >5 reviews where the overall polarity was negative.

## Fairness Determination

To prepare the dataset for model training, we labeled 650 Uber reviews as fair or unfair. We used Uber reviews because they were the most plentiful, and it had an extra predictor for whether the interview was prompted. We considered three factors when determining if a review was unfair: (1) whether the review had a rating < 5 stars [since this analysis is for the benefit of the driver, 5 star reviews are always considered fair], (2) mentioned the driver was fantastic, but lowered the rating because of other reasons, or (3) mentioned things outside of the driver’s control; such as price or excessive time spent waiting for a driver to be matched to the reviewer. 112 out of the 650 reviews were labeled as unfair using this criteria.

# MODELS

## General Information

Since we had a clear idea of the defined classes we wanted to predict, we primarily focused on supervised learning. Numerical data was scaled using sklearn’s MinMaxScaler for quicker training and to avoid certain features having more importance over others for methods that used distance formulas [3]. TF-IDF was used to convert the body and title of the reviews to numerical data and we found that it was best to remove the 50% most common words throughout the dataset from consideration, as these tended to be filler words and decrease the performance of the models. The potential raw inputs for each model were the number of reviews that had been left by that user, the rating of the review, whether the review was asked for by the company (prompted), the review title, and the review body. 5-fold cross validation was used to determine metrics for all models.

The models were instructed to classify a review as fair (0) or unfair (1).

## Multinomial Naïve Bayes (MNB) The Multinomial Naïve Bayes model was appropriate for this task, as it is a popular and effective model known for its strength in text classification. The inputs to this model were simply the review title and a review body (which were then processed using TF-IDF). This model was the most precise but had poor recall and accuracy relative to other models. For this model, the accuracy was 88.62%, the recall was 63.40%, the precision was 69.51% and the F1 was 65.40%. We hypothesized that this was because unfair and fair reviews probably had quite similar content, and it had trouble distinguishing between the two well based solely on probability.

Chart, scatter chart

Description automatically generated

Figure . The classification of each sample in the joint Uber and Lyft dataset by the MNB Classifier (n=6463). The axes were generated by TruncatedSVD due to the original data having several more dimensions.

## Support Vector Classifier (SVC)

Support vector machines are useful for approximating more unusual functions than raw logistic regression, and thus we thought could be useful for classifying these reviews. It received the number of reviews left by the user, the rating, and if the review was prompted as inputs. We used a fourth-degree polynomial for our kernel, as it had the best performance against the test data. The model had good all-around statistics when compared to other models. For this model, the accuracy was 88.15%, the recall was 82.96%, the precision was 62.24% and the F1 was 70.80%. Its weakest metric was precision. This is likely because of the emphasis the rating had in the model - so unfair reviews with higher ratings (generally 3-4 stars) were falsely tagged as fair.

Chart, scatter chart

Description automatically generated

Figure . The classification of each sample in the joint Uber and Lyft dataset by the SVM Classifier (n=6463). The axes were generated by TruncatedSVD due to the original data having several more dimensions.

## Decision Tree Classifier (DTC)

Decision trees are a good solution to this problem due to it leading itself naturally to classification. The DTC received the same input as the SVC model. The Gini impurity was the best method to use to calculate when to split nodes on the tree, and a node would be eligible for splitting if it had at least 32.5% of samples under it. For this model, the accuracy was 88.46%, the recall was 100%, the precision was 60.26% and the F1 was 75.10%. The decision tree classifier had the best recall rate and correctly identified all unfair reviews. However, it was also the least precise, and had many clear-to-human-eyes instances of fair ratings listed as unfair.

Chart, scatter chart

Description automatically generated

Figure . The classification of each sample in the joint Uber and Lyft dataset by the Decision Tree Classifier (n=6463). The axes were generated by TruncatedSVD due to the original data having several more dimensions.

## Voting Classifier

The high precision of the MNB model and the high recall of the DT model resulted in us deciding the best approach was likely to combine the two into one. We did this using a Voting Classifier. The DT classifier was given 70% voting power, while the MNB was given 30% voting power. The Voting Classifier is our final model. For this model, the accuracy was 90.00%, the recall was 80.32%, the precision was 68.41% and the F1 was 73.41%. Though the accuracy is very similar for all models, we believe that is not the primary metric to focus on. This is because there are far more “obviously” fair reviews than unfair reviews and borderline ones, and therefore the number of true negatives will outweigh the other possibilities. The primary focus is on the positives, so a balance between recall (catching as many unfair reviews as possible) and precision (ensuring we are not needlessly flagging fair ratings as unfair) is critical. There is a business case for prioritizing either of the two, but we valued recall. We believed the Voting Classifier was the best model because it is not quite as highly biased towards unfairness as the decision tree classifier, but still had a high F1 score. Based on the validation, it seems to fit the data quite well (see Figure 1).

Chart, scatter chart

Description automatically generated

Figure 11. The classification of each sample in the joint Uber and Lyft dataset by the Voting Classifier (n=6463). The axes were generated by TruncatedSVD due to the original data having several more dimensions.

Table 2. The classification of five cases of interest (not seen before by the model) by the Voting Classifier. The model's decision is given by the "Final Verdict" column, and the "Final Verdict Class Probability" column shows what the probability of selecting that class was.

Graphical user interface, text, application

Description automatically generated

Table . A summary of all the models' performances.

## Table Description automatically generated

## Bisecting-KMeans Model

To experiment, we also tried fitting a Bisecting KMeans  
clustering algorithm to group the reviews into 10 clusters, fed the same inputs as the SVM model. We found that clusters 4 and 5 have the best fair to unfair ratio, but due to the number of unfair reviews that were stuck in “fair” clusters, it was still not suitable to take any action on it. We hypothesized that the reason why this algorithm does so poorly is that “unfairness” is not an obvious quality inherent to the data – it is up to human interpretation. Supervised algorithms can better accommodate this. Increasing the number of features by giving KMeans TF-IDF data and increasing the number of clusters also did not meaningfully help quality. The graph with more clusters is visualized in  
Figure 12.

Chart, scatter chart

Description automatically generated

Figure 12. BisectingKMeans clustering (k=30) for the joint Uber and Lyft dataset (n=6463). The axes were generated  
by TruncatedSVD due to the original data having several more dimensions.

# CONCLUSION

# The original question we sought an answer to was “What factors contribute to unfair ratings in the gig economy and can we detect such metrics to limit their influence on worker’s performance?” We believe we have created decent models to identify these unfair ratings. We have also explored different possible factors that have contributed to unfairness in the reviews, such as the average sentiment of the review date. To explore further, we also grouped together our data in clusters that share attributes that may point to why a certain group of data would be unfair.

# This model could be directly used in the company’s review weighting algorithm, or simply used to quickly flag reviews for review by a human representative. Data analysts and managers for companies in the gig economy can learn a few things from our analysis. Specific days can have negative sentiment, which ultimately may not specifically be because of the drivers, but rather the stress levels on the system for the day or other psychological factors. To explore this further, more data should be used such as app traffic for the day and stress on these apps in specific regions (i.e., New York City) for the date the review was left. Ultimately, it is not the driver’s fault that supply cannot meet demand for at the given instance.

# With more data, we could get more reliable classifications for our data. Ultimately, there is a tradeoff between each of the models we have procured which analysts can adjust to fit what they believe is fair or unfair. With the knowledge of their business rules, they can refine what they believe is an unfair review and train their model to detect such reviews.

# To improve this project, another class representing the “borderline” category could be created to identify those reviews which have valid points and should not be disregarded entirely, but also contain unfair aspects. Unfairness can also be separated into different classes so the reason the review was flagged could be made more obvious. As an example, imagine a “price” class, which identifies reviews made because of the cost of the ride. Clustering algorithms could help in this scenario, when combined with another model to determine unfairness.

# More data and resources aimed at reviewing the feedback and cleaning it should be acquired. More data such as specific region such as city, request traffic, road traffic conditions, and weather conditions can help inform what a valid review may be. Factors such as demand for ride sharing services, the ability for the drivers to deliver on a good service based on their surrounding environment (i.e., weather and traffic conditions) and pricing can all affect the reviews for the driver, yet the drivers are not in control of such factors. Those factors can help assist the analysts in creating an environment that assists on getting feedback for the drivers and make the gig economy a better place for everyone.

# REFERENCES

1. James Cook. 2015. Leaked Charts Show How Uber. *Business Insider*. Retrieved April 25, 2022, from <https://www.businessinsider.com/leaked-charts-show-how-ubers-driver-rating-system-works-2015-2>.
2. TextBlob: Simplified Text Processing — TextBlob 0.16.0. Retrieved April 26, 2023 from <https://textblob.readthedocs.io/en/dev/>.
3. scikit-learn: machine learning in Python — scikit-learn 1.2.2. Retrieved April 26, 2023 from https://scikit-learn.org/.