IDS 705: Final Project - The Tipping Point

Team 1: Peining Yang, Raza Lamb, Michelle Van, Shining Yang

Abstract

As ridesharing companies attempt to pivot to profitability, some research has shown that drivers may make less than minimum wage [1]. This project is motivated to understand user tipping behavior. Specifically, we investigate whether rideshare tipping behavior is predictable and whether meaningful insights about tipping behavior can be extracted. While previous works have investigated this relationship [2], there is little work on whether the problem is better approached using more flexible machine learning methods. The data contains 120,000 trips between November 2018 and January 2020 in Chicago, IL. Four models were fit, including two logistic regressions, a random forest, and XGBoost. The analysis showed that tipping behavior is very individualistic and cannot be reliably predicted from available information, which concurs with previous research [3]. However, some meaningful insights were extracted. Picking up from O'Hare International Airport and driving at noon yields the highest odds of earning a tip.

Introduction

In 2009, Uber introduced the first ridesharing application, allowing users to"hail" a cab from their phones, which led to the rise in ridesharing platforms such as Lyft and Via. They subsequently incorporated a tip feature for riders to tip their drivers on top of the original fare. A research paper [2] analyzed over 40 million Uber trips, and found approximately 15% of rides ended with a tip, while nearly 60% of riders never tip. Given the lowered rate of drivers' base salary, tips can be a crucial aspect of making ridesharing a profitable venture. A natural question for rideshare drivers is thus: "what makes a ride tip-worthy?" Are there strategic decisions drivers can make to increase their probability of receiving a tip? In this project, we analyzed rideshare data for Chicago between 2018 and 2020 to investigate the most important factors that impact tipping. We will also explore the trade-off between interpretability and accuracy, determining whether tips are predictable from available data and whether that prediction is understandable.

Background

Over the years, tipping has become a major contributor and source of income for millions of workers in the US [4, 5]. Different industries such as taxis and restaurants rely heavily on tipping, with the US food industry generating $46.6 billion in tips in 2011 [6, 7]. Tipping constitutes a significant proportion of the US economy, which have garnered a magnitude of scholarly interests in tipping behaviours. Multiple studies have concluded tipping stem from a mix of economic, psychological, sociological and irrational rationales [3, 6, 8–10].

Rideshare apps such as Uber and Lyft have given rise to a surge of new drivers over the past decade — over 233,000 drivers by 2015 in the US, increasing 13% per year [11, 12]. As a result, tipping has become more significant over the years as drivers obtain much of their income from tips. Moreover, the introduction of electronic devices in NY cabs was found to have doubled the tip amount received [13]. Unlike restaurants, drivers do not have much control over the service experience, therefore gestures, tones, conversations are limited [14]. In addition, riders do not typically hire the same driver more than once [15], so the expectation of receiving better future services are not applicable. [16]

Besides service quality and socioeconomic motives to tipping, several studies indicate that weather and customer mood are factors that contribute to tipping. [8] These studies have shown that weather

conditions influence people's mood [17–19], which then can impact tipping. A paper by Devaraj [14] on tipping in taxicab rides in NYC showed a statistically significant positive relationship between sunlight and tipping. Furthermore, tips in Chicago-area restaurants were reported higher during spring. [20]

In March 2020, the U.S. experienced covid-19, which closed down businesses and schools and greatly limited mobility. Research has been conducted [21] that provides background on the tipping trend during the pandemic, and Lynn [22] reported that the average tip of pizza delivery drivers increased from January 2020 to July 2020. Although this research showed a general trend on the magnitude of tips, it did not provide insights on the likelihood of tipping.

Ridesharing apps currently do not have their data accessible to the public. However, in 2019, Chicago was the first city to mandate open data on all ride-hailing services on all trips including pick-up/drop-off locations and trip fare tip [23]. Using this information, we conducted research on predicting whether the driver gets tipped based on multiple factors including but not limited to weather conditions, geospatial location, and other ride features. Most of the research on tipping have been carried out mainly on restaurants [4,9,24,25], and the few on ridesharing and taxi services were typically based on surveys due to limited data. The small number of research papers on tipping and taxi services have been conducted in New York [3, 16, 26], while there is only one paper on the effect of the pandemic on tipping in Chicago [21]. This report utilized taxi data from Chicago, and concluded that the likelihood a tip was left to taxi drivers went down by 5%, but the magnitude of the tip increased by 2%. Here, we expand on previous work by extending from taxis to ridesharing service and utilizing more flexible machine learning methods to evaluate the predictability of tipping.

Data

In April 2019, Chicago required all ride-sharing companies to report data on rides beginning or ending within city limits. This data [27] is available publicly, and contains the following information:

- Trip start time and end time (rounded to nearest 15 minutes)
- Trip length (miles)
- Trip duration (seconds)
- Pickup and drop-off location (census tract)
- Fare (rounded to nearest $2.50)
- Tip (rounded to nearest $1)

As of April 9th, 2022, the data contains 239,898,489 rides. We used the Socrata Open Data API to extract the data directly from its source. We utilize data from November 2018-January 2020 for training and April-July of 2020 for testing. To further reduce the data to a workable size, we randomly sampled 0.1% of the data, stratified by day. In addition to the rideshare data, we also included weather data, extracted from the National Oceanic and Atmospheric Administration (NOAA) API. The final data set contains 119,576 observations.

Exploratory Data Analysis

We first explored the distribution of our outcome variable, whether a rider tips or not. From Figure 1, there is a significant imbalance, only 17.86% of riders tipped. We then explored the relationship between our outcome variable and potentially significant predictors. Intuitively, we expected that with an increase in trip mileage, the rider would be more likely to tip. On average, trips with tips span 5.47 miles while trips without span 4.72 miles. This trend is also reflective in trip duration, as

trips with tips, on average, last longer than those without. In addition, we explored the impact of pickup/drop-off locations on tips. Figure 2 contains a map of proportion of trips ending with tips by the pickup community area. There are two community areas with relatively larger proportion of rides with tips, which are O'Hare and Garfield Ridge. The southern region of Chicago generally tip less than others. We observe a similar trend with dropoff locations.

Figure 1: Class Imbalance Distribution

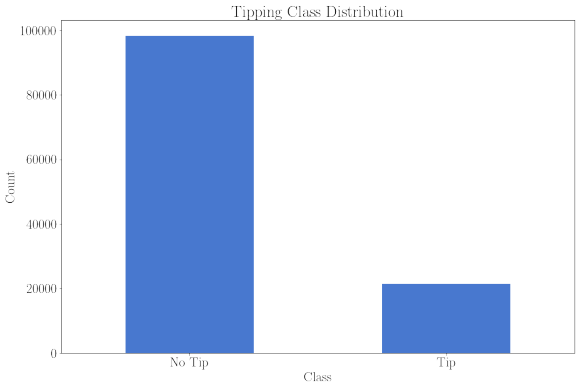
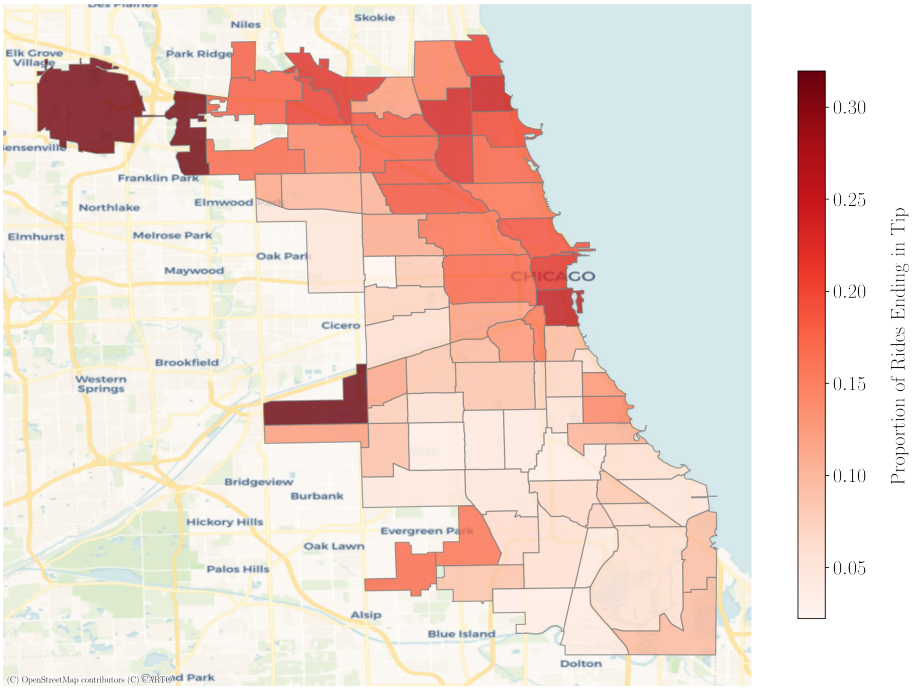
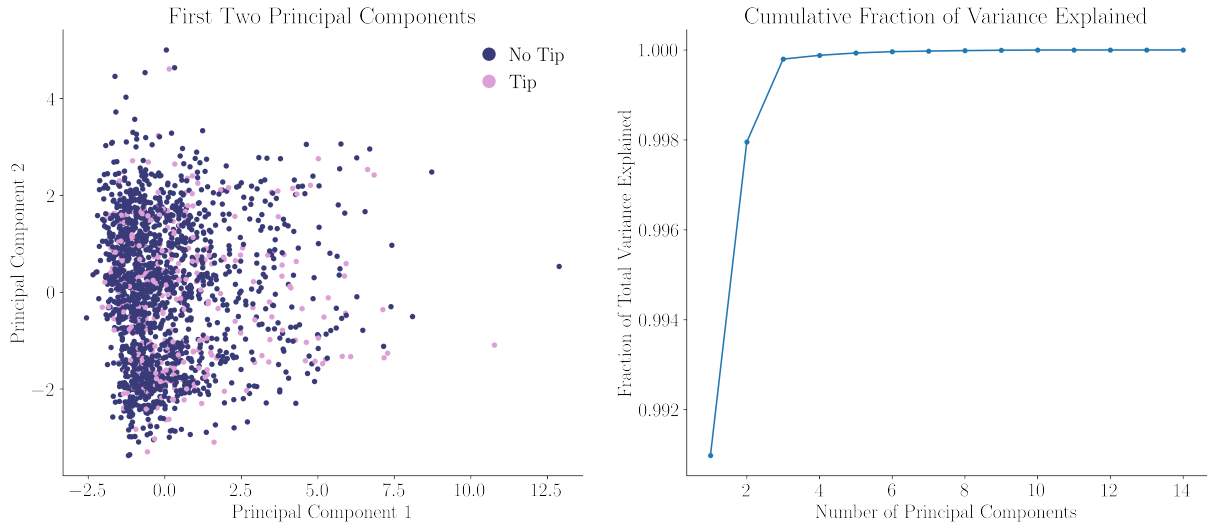


Figure 2: Proportion of Rides Ending in Tip by Pickup Area



Additionally, we performed dimension reduction using Principal Component Analysis (PCA) for visualization. Figure 3 displays that there are no discernible cluster patterns between tip and no tip when visualized in two-dimensions. Additionally, around 99.8% of the variance is explained by the first two principal components.

Figure 3: Visualization of Dimension Reduction.



Methods

Figure 4 depicts the process of how we conducted our experiments.

1. Data Preparation

1.1 Preprocessing

Data with missing pickup/drop-off locations indicate that the trip started or ended outside Chicago. We removed these observations to focus our analysis on trips completed in the city. There were several extreme outliers in our data that were evident in our EDA – these were removed.

Due to the imbalance found during EDA, multiple oversampling and undersampling methods were performed [28]. Although these resampling techniques balanced our dataset, issues such as overfitting were met when performing modeling, which consequently led to less accurate models and poor generalization to the validation set. Therefore, we proceeded without balancing the dataset.

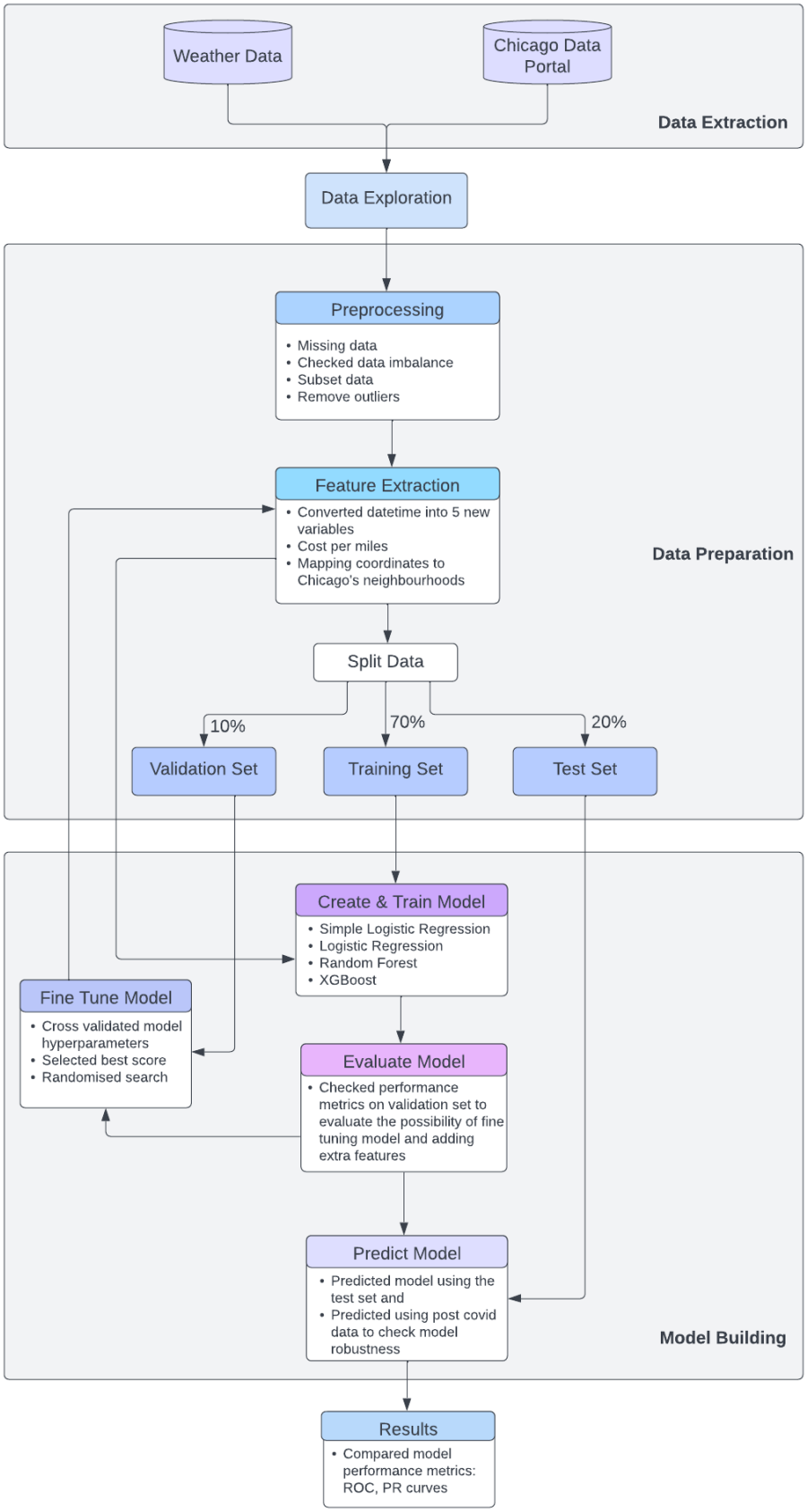
1.2 Feature Extraction

Various feature engineering methods were performed to transform our raw data into representative features. These included:

- Weekend indicator
- Public (Federal) holiday indicator
- Categorical season variable
- Categorical time of day variable - e.g., early morning, midday, morning rush hour

We also extracted the cost per mile by dividing the total cost by the total miles travelled by the driver. In addition, the pickup and drop off locations, composed of latitude and longitudes co-ordinates, were converted into Chicago's neighbourhoods [29] to enable representation of the drop-off and pickup locations.

Figure 4: Experimental Design Flowchart



2. Model Selection

We utilized three models: Logistic Regression, Random Forest Classifier and XGBoost Classifier. The data was split into training, test, and validation set with a 70%, 20% and 10% proportion. Due to the imbalance issues in our data, accuracy was not an appropriate evaluation metric. AUC (Area Under the Receiver Operating Characteristic Curve) was chosen instead, as this provides a better indication of overall performance. AUC was used as the criteria in both hyperparameter tuning and overall performance evaluation.

2.1 Logistic Regression

Two logistic regression models were fit to the the data. First a baseline model that included only variables that a driver could perceive while driving, and then a second model that included all variables. Logistic regression was chosen as the baseline because it is less inclined to over-fitting and the predicted parameters gave inference about the importance of each feature.

For the second logistic regression model, we investigated both L1 and L2 regularization, tuning the regularization strength. In this case, the models performed worse under regularization, so it was abandoned and the final model was fit without regularization.

2.2 Random Forest

We chose Random Forest in our experiment because it is more accurate than simple decision trees and is less likely to have overfitting issues. We utilized Randomized Search to fine-tune our model. Parameters searched include an: number of trees in the forest, maximum depth of the trees, minimum number of samples required to split an internal node, minimum number of samples required to be at a leaf node, and number of features to consider when looking for the best split.

2.3 XGBoost

XGBoost (Extreme Gradient Boosting) was chosen because not only does it provide parallel tree boosting and enhanced computational efficiency, it also utilized information learned from previous models to fit the next model to promote better performance. While Random Forest's "bagging" method minimizes the variance and overfitting, XGBoost's "boosting" method minimizes the bias and underfitting.

Again, Randomized Search method was used to fine-tune our model. The hyperparameters included in the tuning process were: learning rate, minimum loss reduction required to make a split, and maximum depth of a tree.

Results

To evaluate the models, we investigated the performance of our models on predicting tipping behavior and the interpretability of the models for extracting meaningful insights.

1. Performance

All four models were evaluated on held-out test set from the period on which they were trained. The ROC curves and Precision/Recall (PR) curves are shown in in Figure 5 and Figure 6 respectively. As is visible from the AUC and average precision, XGBoost model performed the best, but all models did not achieve high performance.

Figure 5: ROC Curves for Four Final Models on Pre-covid Data

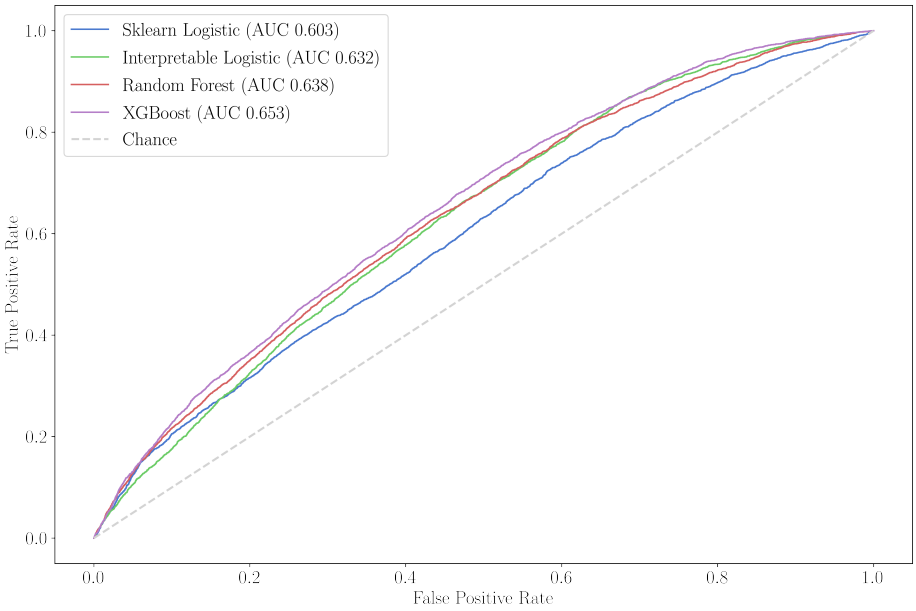
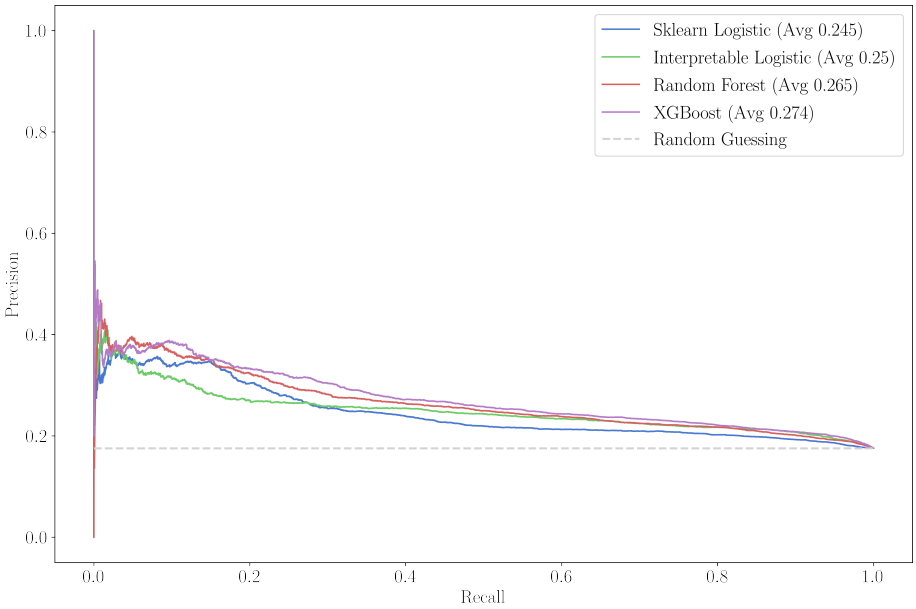


Figure 6: Precision/Recall Curves for Four Final Models on Pre-covid Data



We also evaluated generalization performance in a new time period. Specifically, we utilized data from April through July 2020. Theoretically, this new data was a strong test of generalization, because research suggested that tipping behavior did change during the pandemic [21]. The ROC and PR curves are included here for this data in Figure 7 and Figure 8. Surprisingly, the models performed similarly on data from this time period. In fact, the simple logistic regression performed noticeably better on this data than on the original test data. This indicated that despite relatively poor performance, our models were quite robust.

Figure 7: ROC Curves for Four Final Models on Post-covid Data

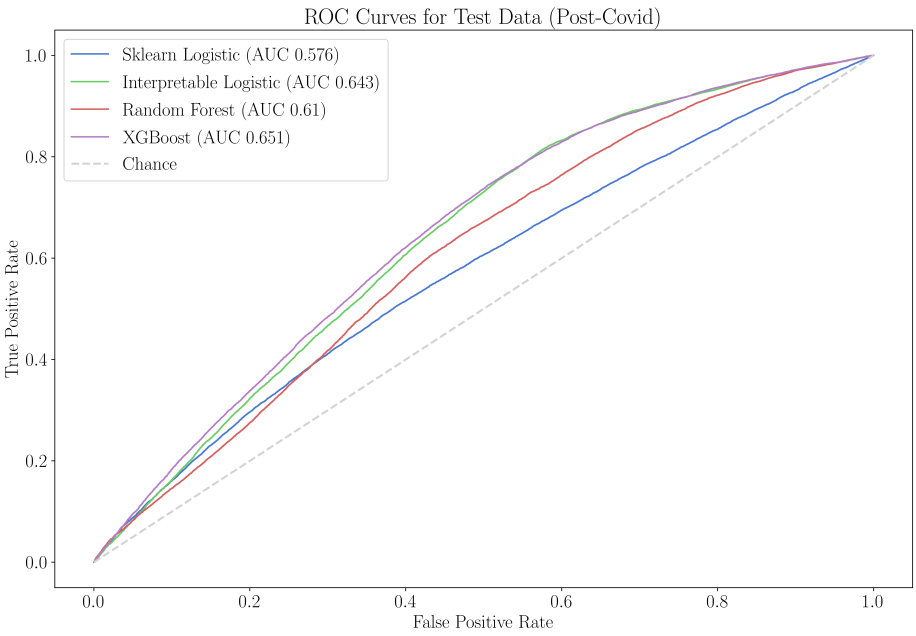
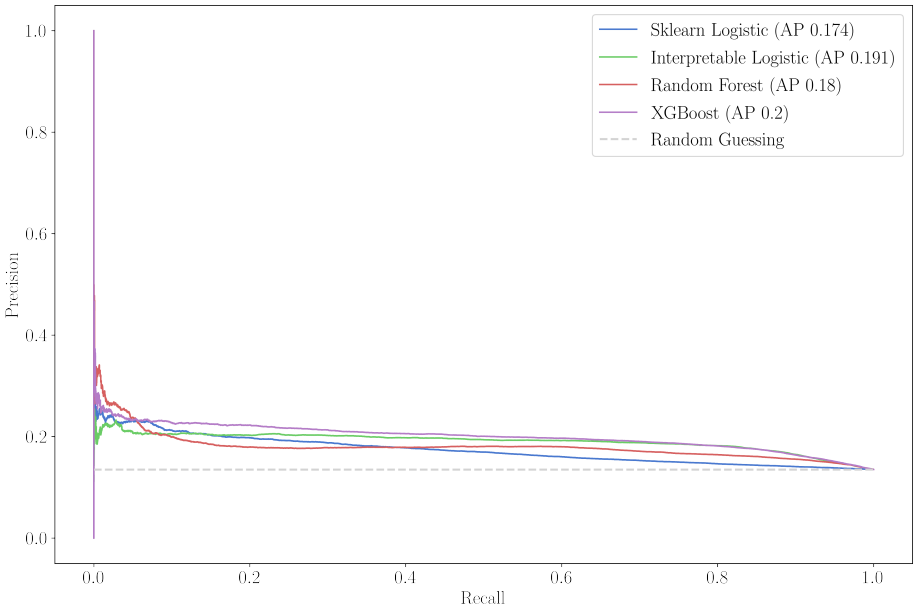


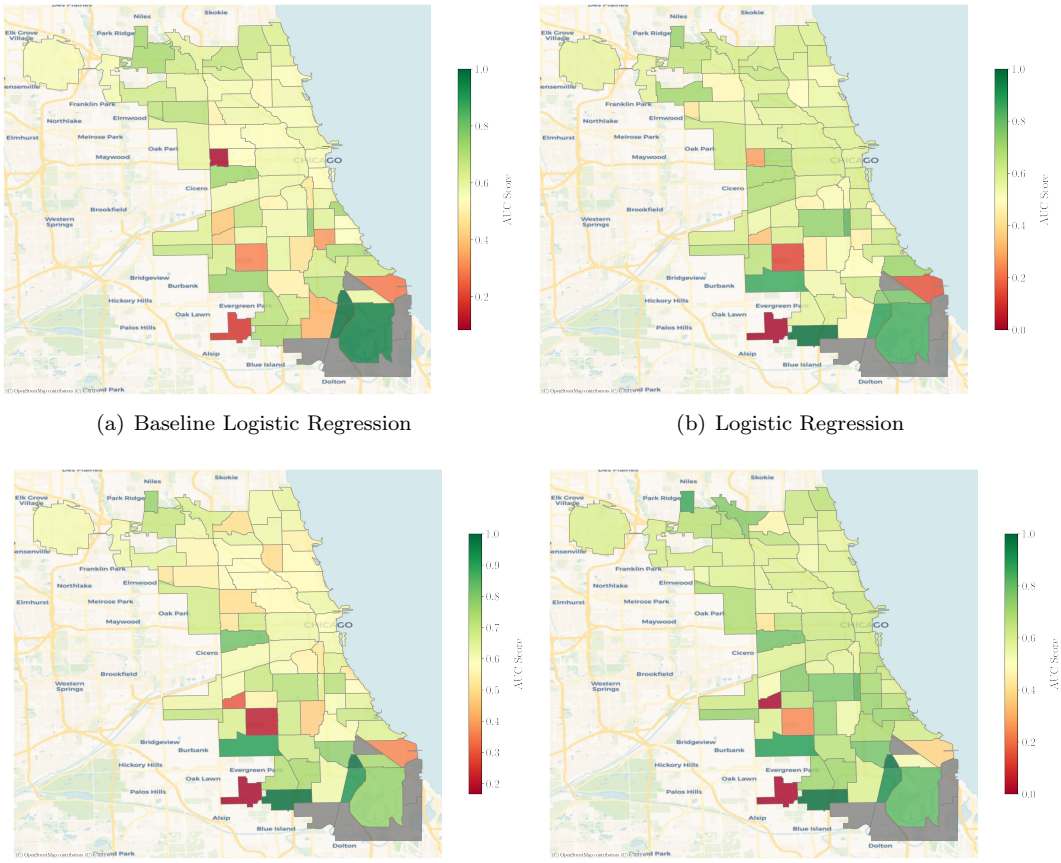
Figure 8: Precision/Recall Curves for Four Final Models on Post-covid Data



We were also interested in spatial performance and potential bias in the models. To evaluate this, we calculated AUC separately for each model and each pickup community area. The results are displayed in Figure 9. The grey areas indicate community areas from which the test data did not contain any tippers. We found that there is spatial correlation in the performance of all the models, with the neighborhoods in the southern portions of Chicago having greater variability (and generally worse performance). This indicated that there could be bias in our model, especially considering that the South side of Chicago has significantly higher Black and Hispanic populations, compared

to downtown and the North side.

Figure 9: AUC of ROC Curve by Community Pickup Area



(a) Baseline Logistic Regression

(b) Logistic Regression

(c) Random Forest

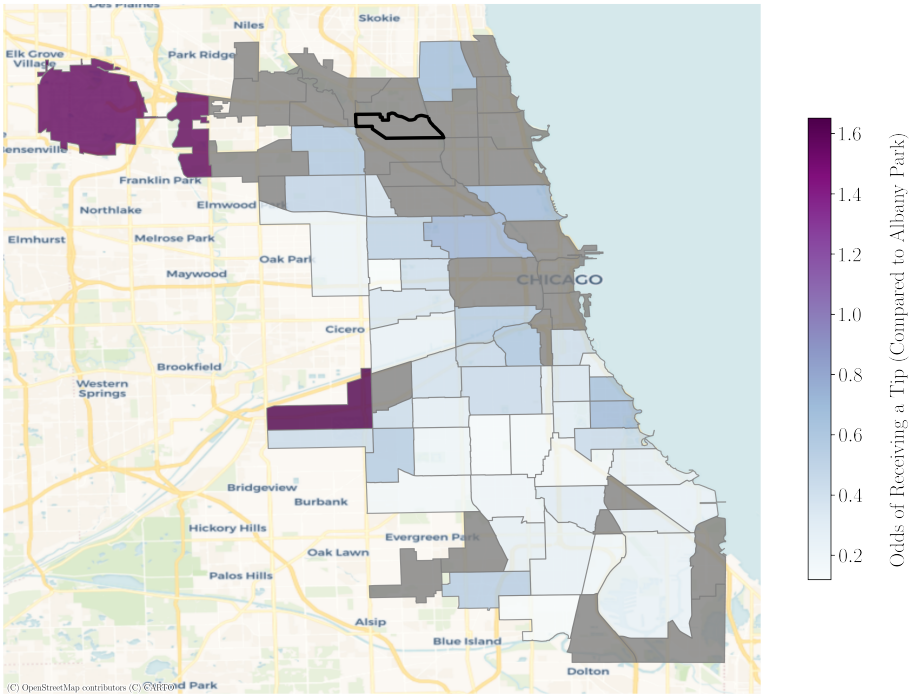
(d) XGBoost

2. Interpretability

While performance is always a desirable goal, in this specific case, the predictions are only useful if there are meaningful insights attached to them. Here, we briefly discuss the interpretability of three of our four models. Unfortunately, the logistic regression conducted in scikit-learn was not interpretable because it only provided the coefficients, and not standard errors or confidence intervals.

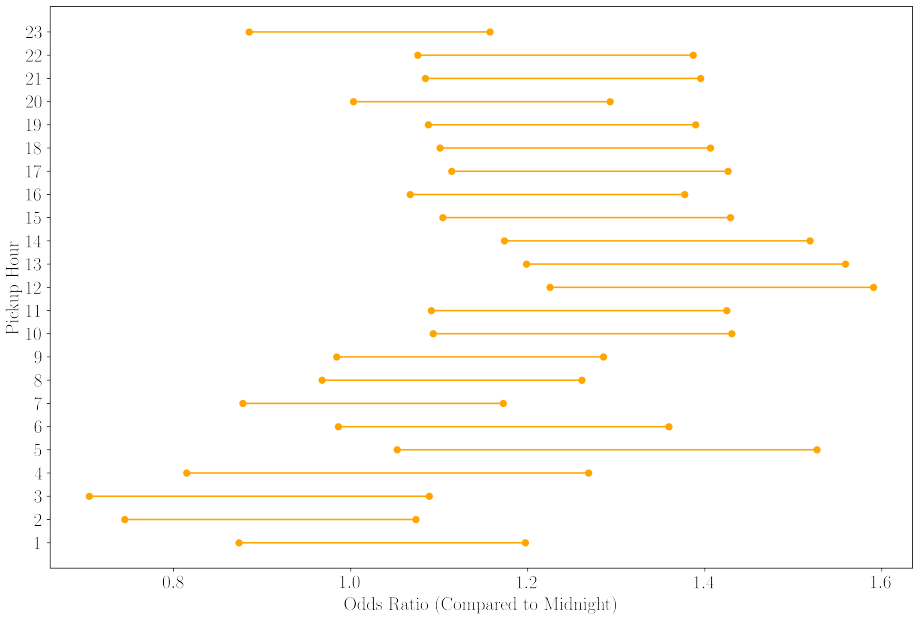
In the case of the simple logistic regression, the results were inherently interpretable. The coefficients and standard errors were used to determine odds ratios. Figure 10 displays the odds ratio of receiving a tip compared to Albany Park (outlined in black). Grey neighborhoods represent areas in which the 95% confidence interval of the odds ratio includes 1 (indicating no difference from Albany Park). Clearly, there is a significant amount of spatial correlation, with neighborhoods near each other experiencing similar odds of tipping. One notable standout, also observed in EDA, is O'Hare international airport, with a significantly higher relative odds of tipping. Another notable predictor in the data included whether or not the ride was authorized for a"shared-trip", indicating that the rider may share their ride with other unknown passengers. This simple indicator was a strong negative predictor of tipping, with rides authorized for a "shared-trip" having a significantly lower odds of receiving a tip.

Figure 10: Odds Ratio of Receiving a Tip by Pickup Area.



Continually, there also appear to be time effects on tipping, as demonstrated in Figure 11. Compared to midnight, midday (12:00 - 3:00 pm) appears to yield larger likelihood of tipping, while early morning (1:00 - 3:00 am) riders are less likely to tip. Clearly, there are significant insights we could extract from the simple logistic regression model.

Figure 11: Odds Ratio of Receiving a Tip by Pickup Hour.



For Random Forest and XGBoost classifiers, we extracted the top 10 important features from both of them, included below in Figure 12 and Figure 13. In both models, the additional charges encountered in a ride and whether or not the ride is a"shared trip" are the most important features. However, the rest eight features varied. This is because the feature selection process in Random Forest and XGBoost is different - when there is high correlation between features, XGBoost will pick one feature to process while ignoring some/all of the remaining correlated features; but in Random Forest, the tree is not built on certain features, instead the features are selected randomly at first, and then the model would learn different correlations of different features. While this is interesting, it does not provide meaningful insights to rideshare drivers, and cannot be utilized directly to see how these factors influence tipping behavior.

Figure 12: Top 10 Important Features for Random Forest Model.

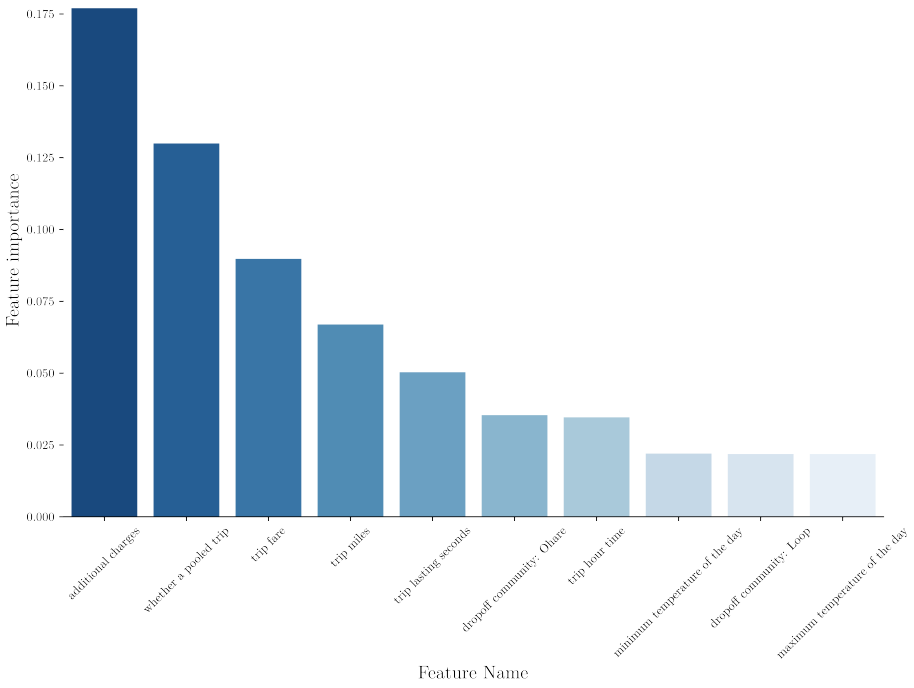
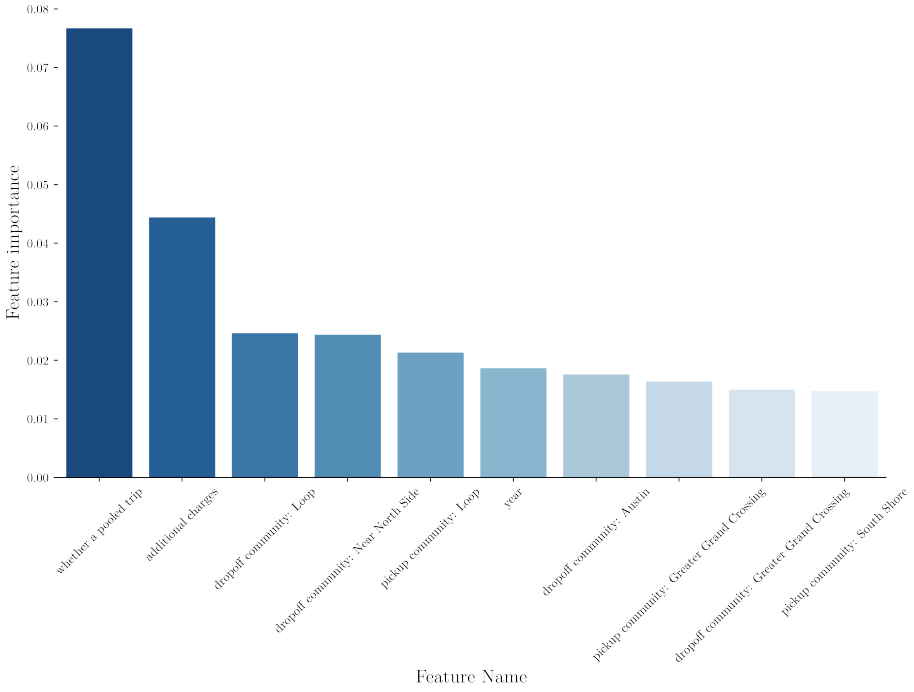


Figure 13: Top 10 Important Features for XGBoost Model.



Conclusions

Overall, the baseline logistic regression model appears to be the ideal model in this space. Machine learning methods with a more flexible form only marginally increased performance, while dramatically decreasing explainability and interpretation. The findings are consistent with previous research that suggests tipping behavior is difficult to predict, while extending the analysis to a new functional form (ride-share). Also, despite the pandemic, the models predicative ability remained steady during this new time frame.

However, there are some key limitations to this study. By nature, the analysis is only relevant to rides within Chicago, and even then only rides that begin and end within city boundaries. While only using 0.1% of the total available data made analysis feasible, it also may have limited the insights available. Future work in this space could include a cloud-computing based approach to attempt a larger data sample. Continually, combination of rideshare data with survey data about passenger and driver characteristics could provide better insights.

Roles

Raza Lamb: Responsible for data merging, cleaning, and EDA. Also responsible for the simple logistic regression, plotting, spatial analysis, and writing the abstract and results section.

Michelle Van: In terms of coding, Michelle was responsible for the data preprocessing, feature engineering and logistic regression modelling. For the report, Michelle was responsible for writing the background, methods (pre-processing, feature extraction + flowchart diagram, and logistic regression).

Peining Yang: Responsible for data extraction from the Chicago Data Portal using Socrata Open Data API and presentation. Peining was also responsible for writing the introduction, data and

limitations sections of the report.

Shining Yang: In terms of coding, Shining did data preprocessing; Random Forest and XGBoost model building, tuning and predicting; generating AUC-by-community maps for both models. In terms of the report, Shining wrote contents related to Random Forest and XGBoost in Method and Result sections.

References

[1] J. A. Parrott and M. Reich, "A minimum compensation standard for seattle tnc drivers," *https://irle.berkeley.edu/files/2020/07/Parrott-Reich-Seattle-Report_July − 2020.pdf*, 2020.

[2] e. a. Chandar, Bharat, "The drivers of social preferences: Evidence from a nationwide tipping field experiment," *National Bureau of Economic Research*, 2019.

[3] D. Elliott, M. Tomasini, M. A. C. Oliveira, and R. P. de Menezes, "Tippers and stiffers: An analysis of tipping behavior in taxi trips," *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–8, 2017.

[4] O. Azar, "What sustains social norms and how they evolve? the case of tipping," *Journal of Economic Behavior Organization*, vol. 54, pp. 49–64, 05 2004.

[5] J. S. Seiter, "Ingratiation and gratuity: The effect of complimenting customers on tipping behavior in restaurants.," *Journal of Applied Social Psychology*, vol. 37(3), p. 478–485, 2007.

[6] O. H. Azar, "Business strategy and the social norm of tipping," *Journal of Economic Psychology*, vol. 32, no. 3, pp. 515–525, 2011.

[7] M. Conlin, M. Lynn, and T. O'Donoghue, "The norm of restaurant tipping," *Journal of Economic Behavior Organization*, vol. 52, no. 3, pp. 297–321, 2003.

[8] M. Lynn and M. McCall, "Gratitude and gratuity: a meta-analysis of research on the service-tipping relationship," *The Journal of Socio-Economics*, vol. 29, no. 2, pp. 203–214, 2000.

[9] O. H. Azar, "Tipping motivations and behavior in the u.s. and israel," *Journal of Applied Social Psychology*, vol. 40, no. 2, pp. 421–457, 2010.

[10] O. H. Azar, "The economics of tipping," *Journal of Economic Perspectives*, vol. 34, pp. 215–36, May 2020.

[11] B. of labor statistics, "taxi drivers and chauffeurs," *u.s. department of labor, occupational outlook handbook*, vol. 2016-2017 edition, pp. Available: https://www.bls.gov/ooh/ transportation–and–material–moving/taxi–drivers–and–chauffeurs.html, March 2017.

[12] M. Lynn, G. M. Zinkhan, and J. Harris, "Consumer Tipping: A Cross-Country Study," *Journal of Consumer Research*, vol. 20, pp. 478–488, 12 1993.

[13] M. Grynbaum, "In new york, taxi revenue and tips from credit cards rise.," *The New York Times.*, vol. 2016-2017 edition, 2009.

[14] S. Devaraj and P. C. Patel, "Taxicab tipping and sunlight," *PLOS ONE*, vol. 12, pp. 1–16, 06 2017.

[15] D. Flath, ""why do we tip taxicab drivers?"," *Japanese Economy*, vol. 39, pp. 69–76, Oct 2012.

[16] K. B. Donkor, "How difficult is tipping? nonparametric and parametric estimates of decision costs," 2020.

[17] M. R. Cunningham, "Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan.," *Journal of Personality and Social Psychology*, vol. 37(11), 1947-1956.

[18] P. L. v. A. M. Denissen JJ, Butalid L, "The effects of weather on daily mood: a multilevel approach," *Emotion.*, vol. 8(5), 05 2008.

[19] S. G. Saunders and M. Lynn, "Why tip? an empirical test of motivations for tipping car guards," *Journal of Economic Psychology*, vol. 31, no. 1, pp. 106–113, 2010.

[20] R. B. Cialdini, "Descriptive social norms as underappreciated sources of social control," *Psychometrika*, vol. 72, no. 2, p. 263, 2007.

[21] S. Conlisk, "Tipping in crises: Evidence from chicago taxi passengers during covid-19," *Journal of Economic Psychology*, vol. 89, p. 102475, 2022.

[22] M. Lynn, "Did the covid-19 pandemic dampen americans' tipping for food services? insights from two studies," *Compensation & Benefits Review*, vol. 53, no. 3, pp. 130–143, 2021.

[23] C. M. A. for Planning., "New data allows an initial look at ride hailing in chicago," pp. ́https://data.cityofchicago.org/Transportation/Transportation–Network–Providers–Trips/m6dm–c72p/data; 2010.

[24] M. Lynn, "Service gratuities and tipping: A motivational framework," *Journal of Economic Psychology*, vol. 46, 12 2014.

[25] O. Azar, "The implications of tipping for economics and management," *International Journal of Social Economics*, vol. 30, 10 2003.

[26] C. Antoniades, D. Fadavi, and A. F. Amon, "Fare and duration prediction : A study of new york city taxi rides,"

[27] O. Azar, "Transportation network providers-trips," *Chicago Data Portal.*, pp. https://data.cityofchicago.org/Transportation/Transportation–Network–Providers–Trips/m6dm–c72p/data, 2018.

[28] J. R. R. Mohammed and M. Abdullah., "Machine learning with oversampling and undersampling techniques: Overview study and experimental result.," *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–248, 2020.

[29] C. M. A. for Planning., "New data allows an initial look at ride hailing in chicago," pp. ́https://data.cityofchicago.org/Facilities–Geographic–Boundaries/Boundaries–Community–Areas–current–/cauq–8yn6; 2010.