

Predicting Check-ins of Foursquare Users in Tokyo

Springboard Data Science Career Track: Capstone Project 2

Kevin Limkailassiri

Adviser: Jan Zikeš

Motivation

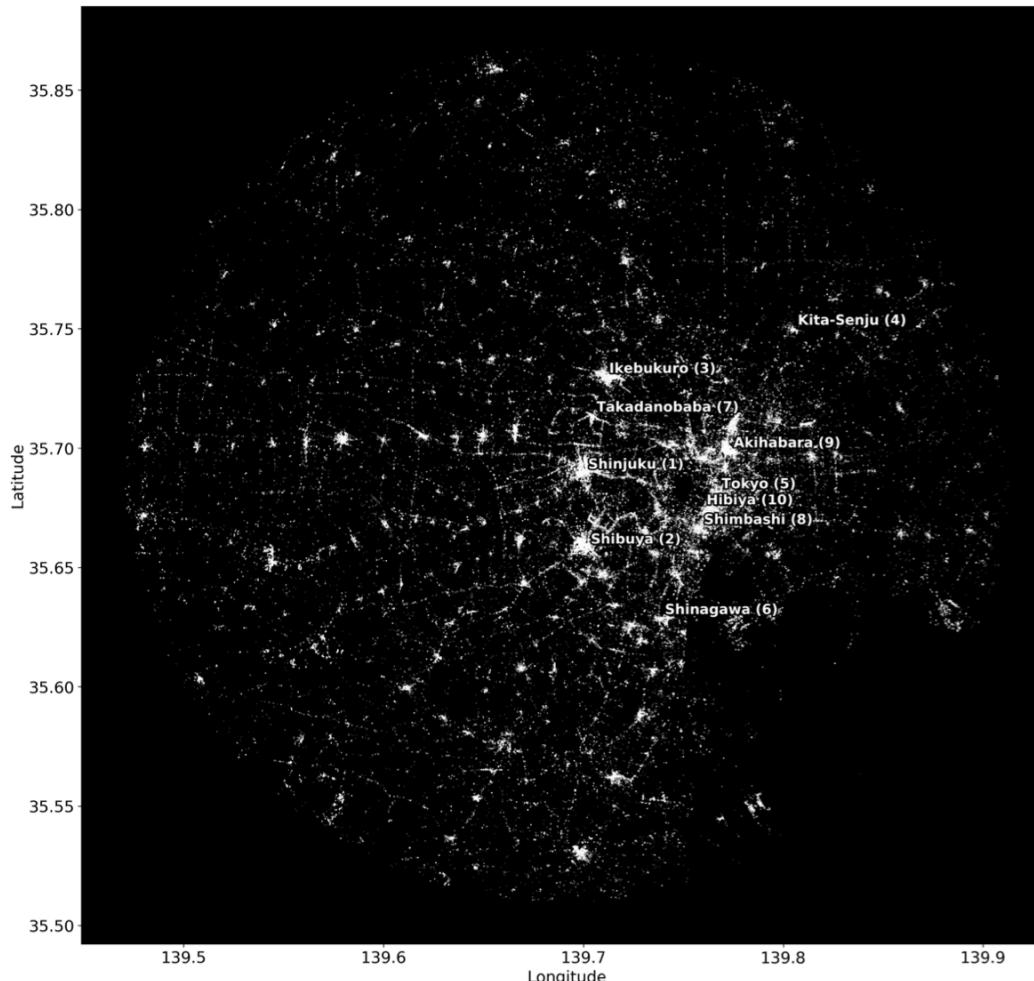
- Foursquare's slogan:

*“Foursquare helps you find the places you’ll love,
anywhere in the world.”*

- What if we could predict what venue a user will check into?
 - Foursquare users win: discover new places they haven’t visited
 - Businesses win: gear up promotions when users would most likely check-in

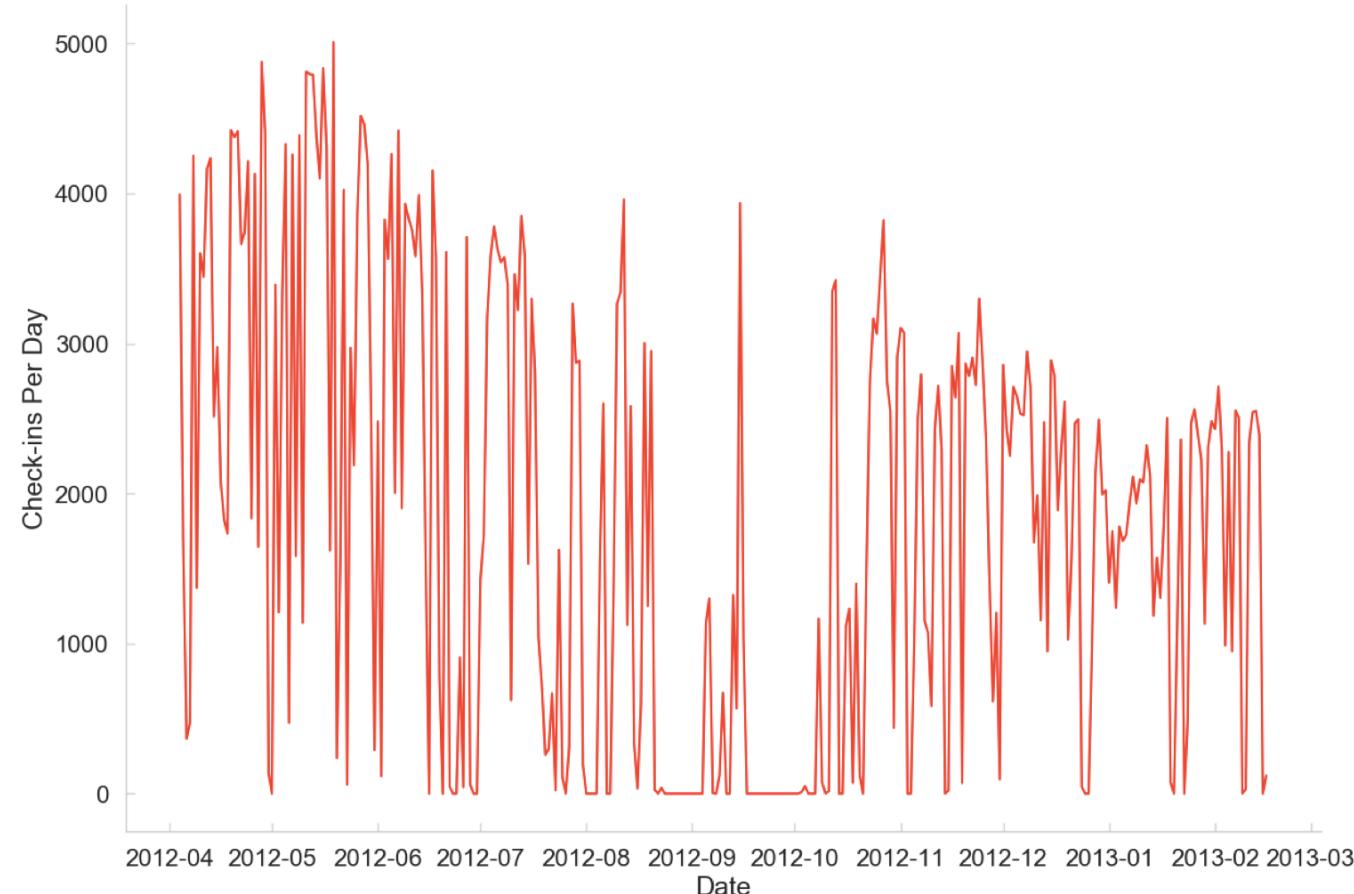
Tokyo: A Foursquare city

- Check-ins are concentrated near train stations and subway stations.



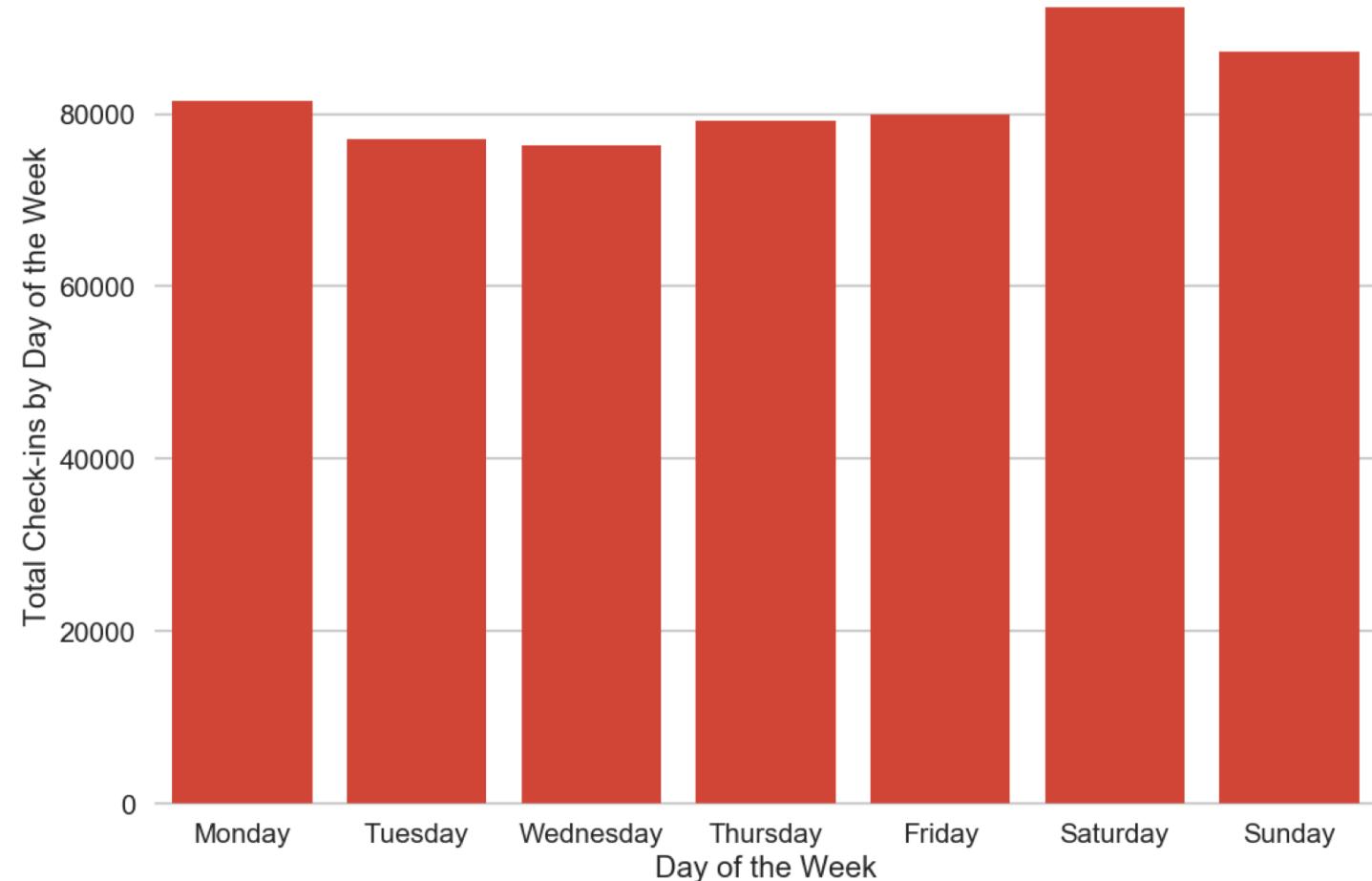
Check-ins per Day During a 10-Month Period

- Missing check-ins will affect monthly averages.
- Trends will be studied on a weekly or daily basis.

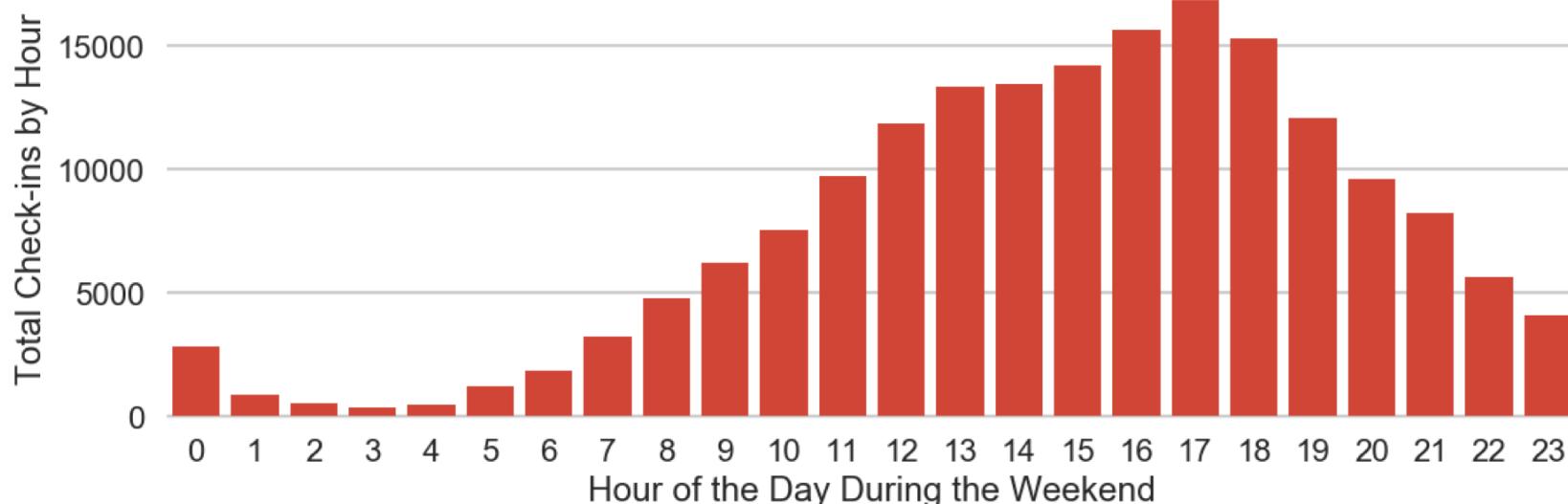
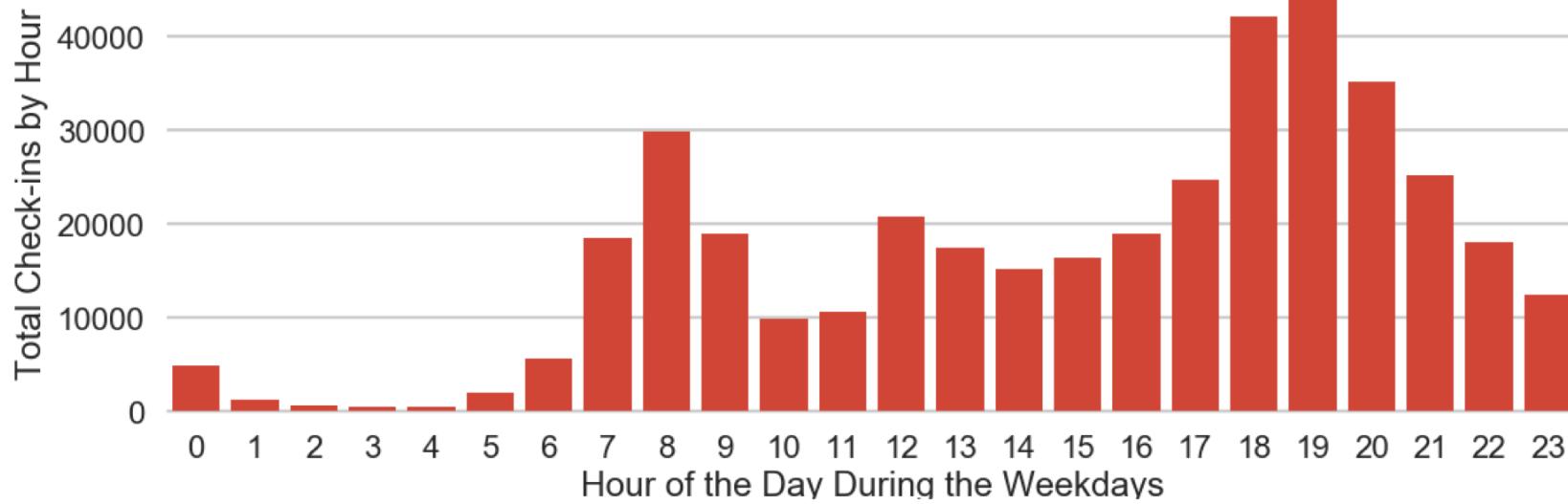


Check-ins by Day of the Week

- Check-ins peak during the weekends



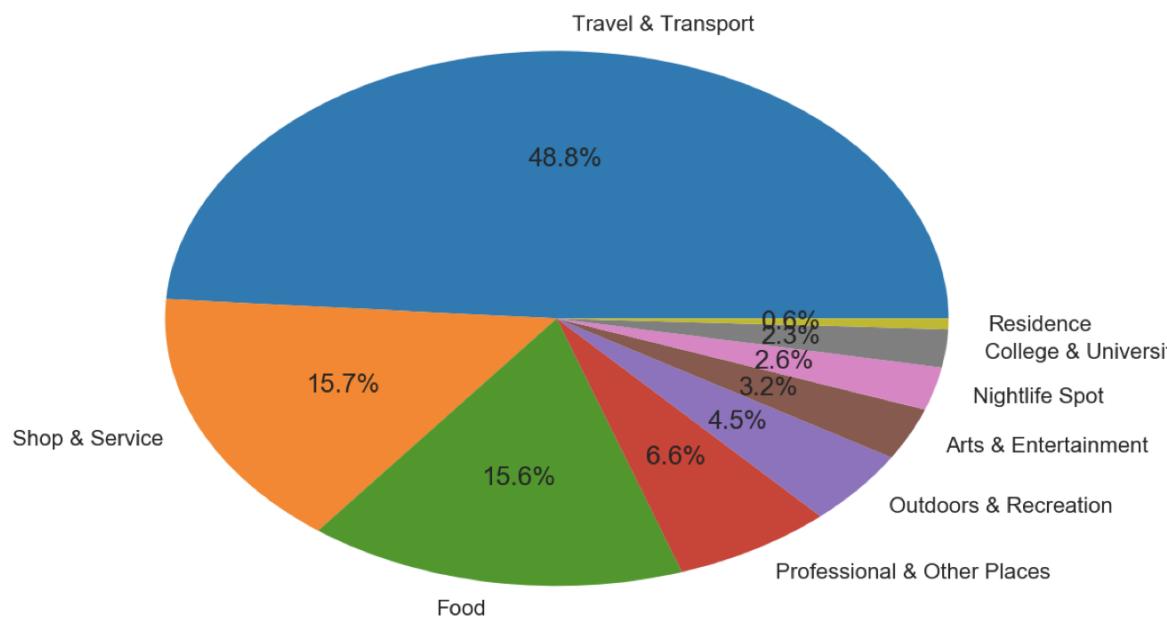
Check-in Activity is Different for Weekdays and Weekends



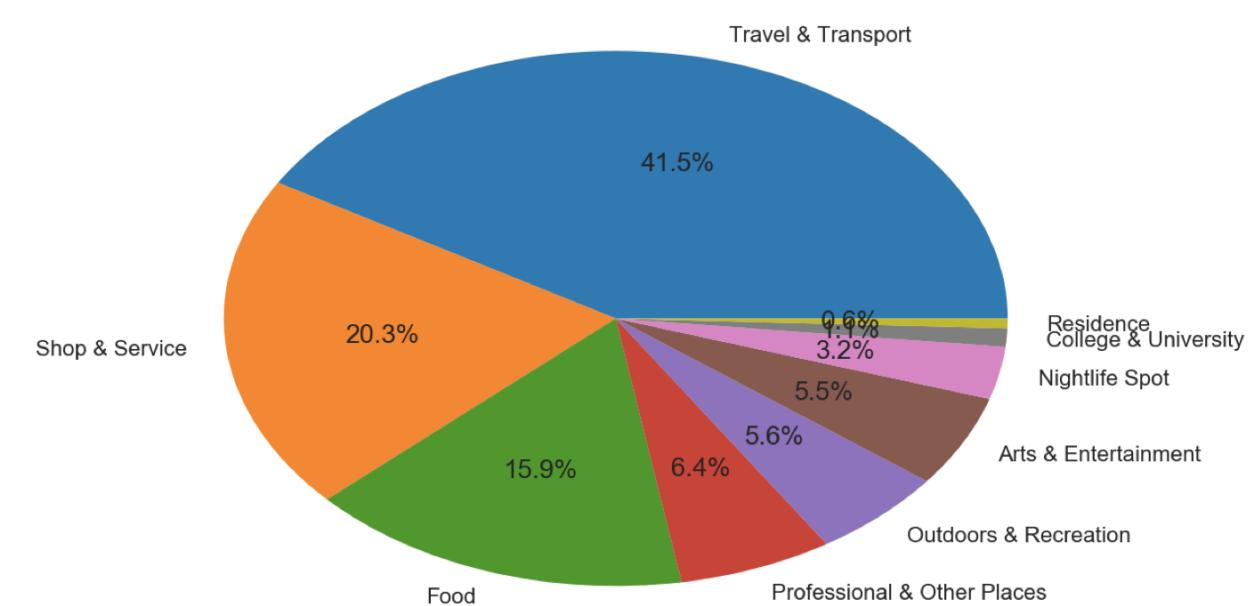
What Kinds of Venues Do Users Check Into?

- The data set is imbalanced with a majority of check-ins for Travel & Transport.

Percentage of Weekday Check-ins For Each Venue Category

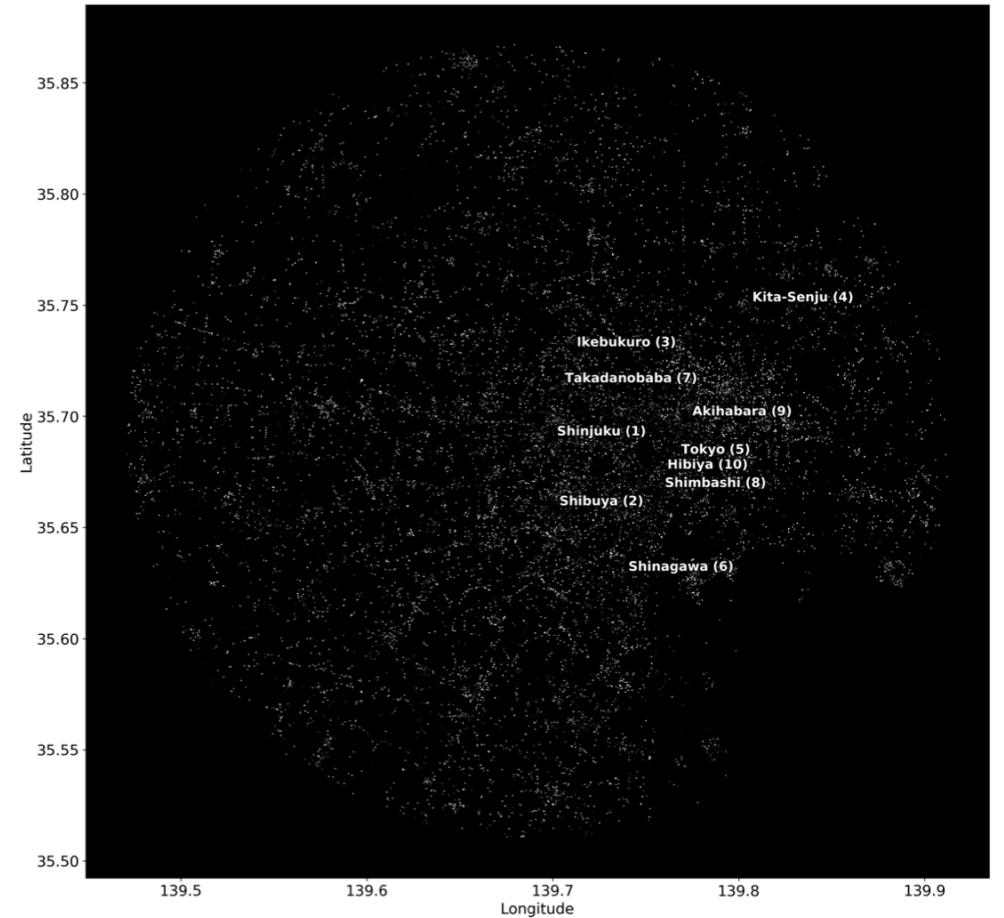
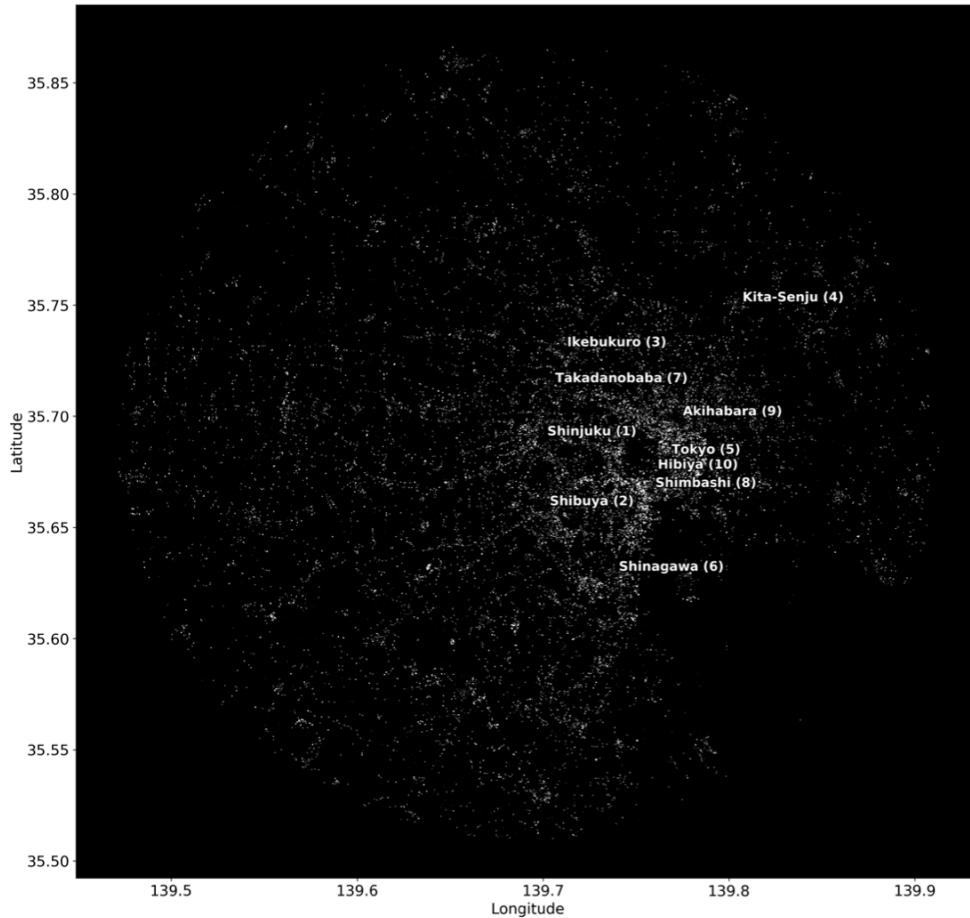


Percentage of Weekend Check-ins For Each Venue Category



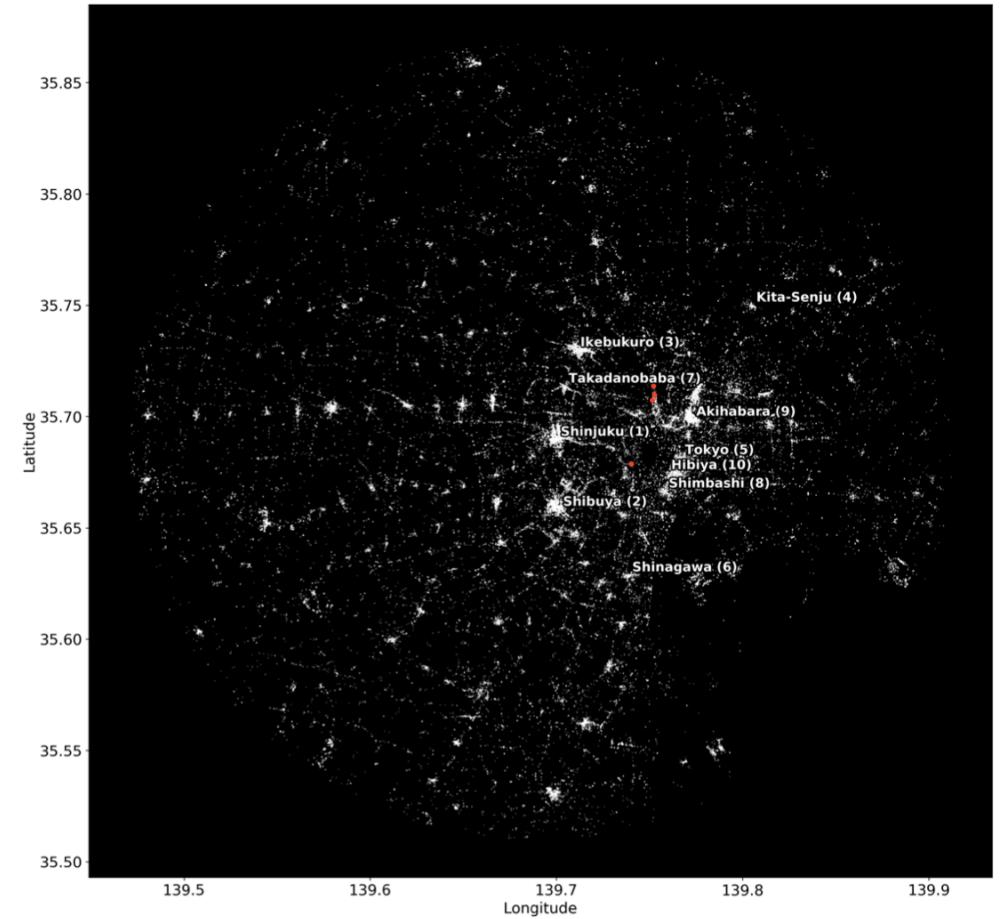
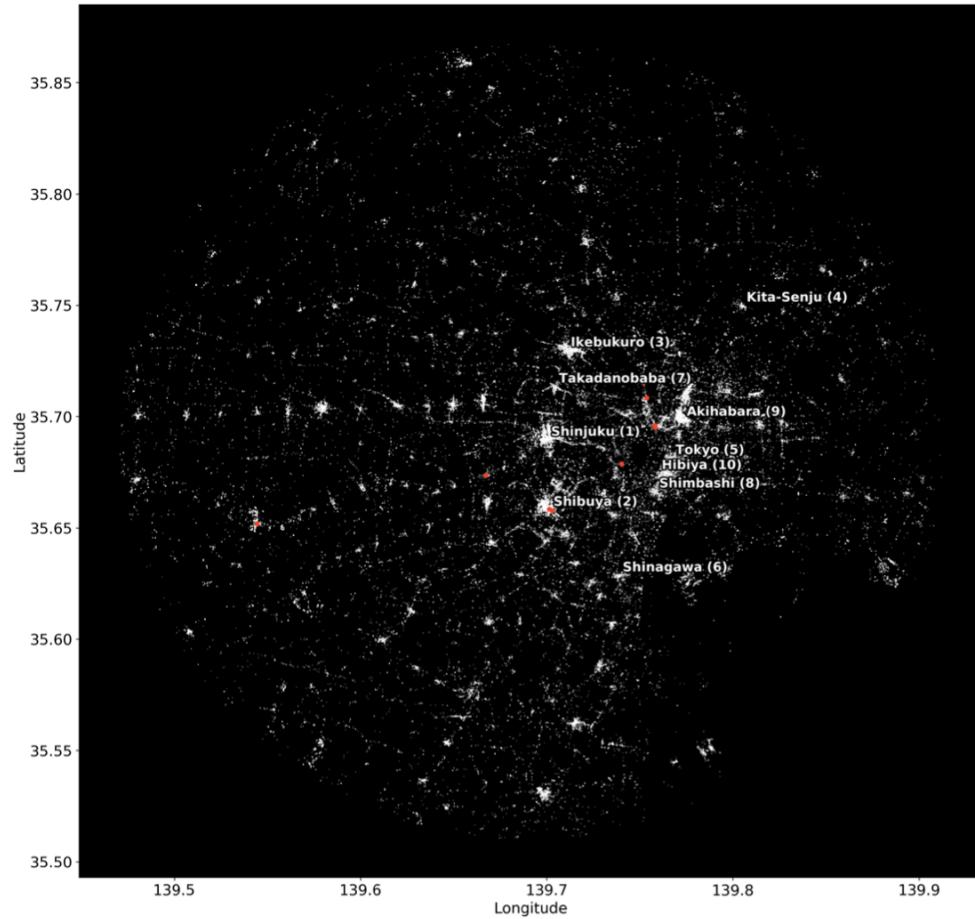
Where People Go On Weekdays and Weekends

- Weekday check-ins are centralized; weekend check-ins are not.



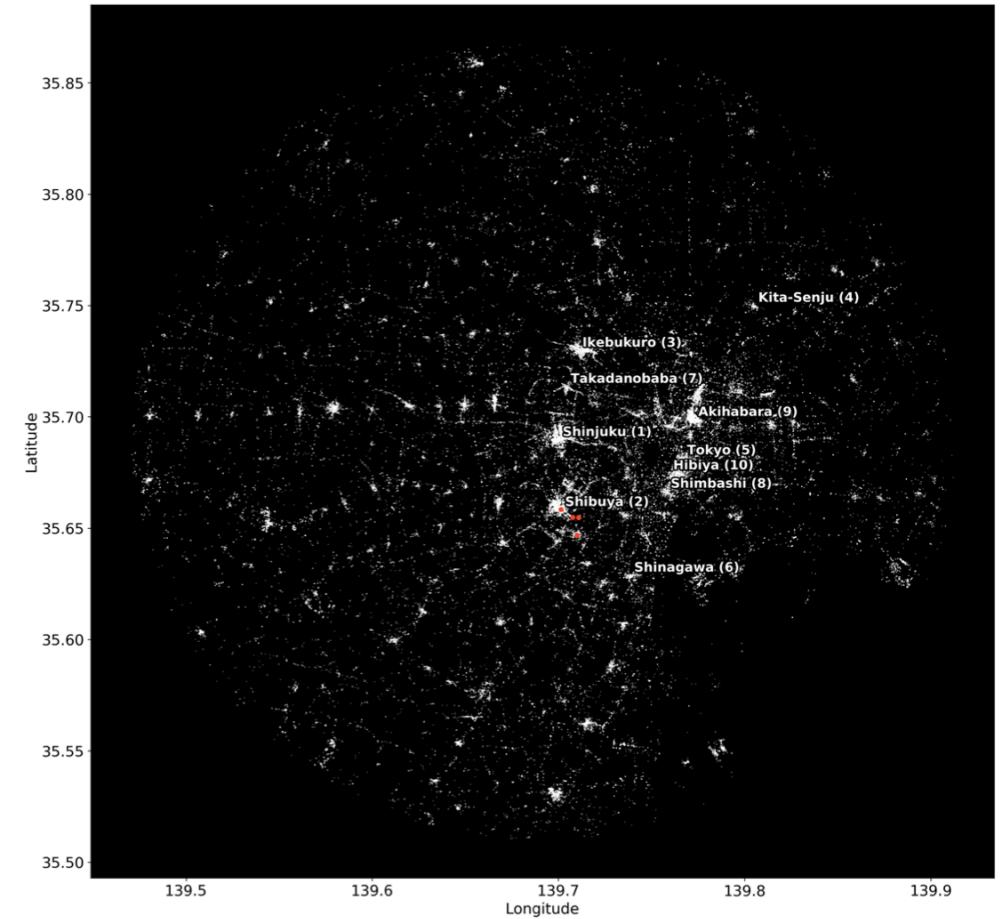
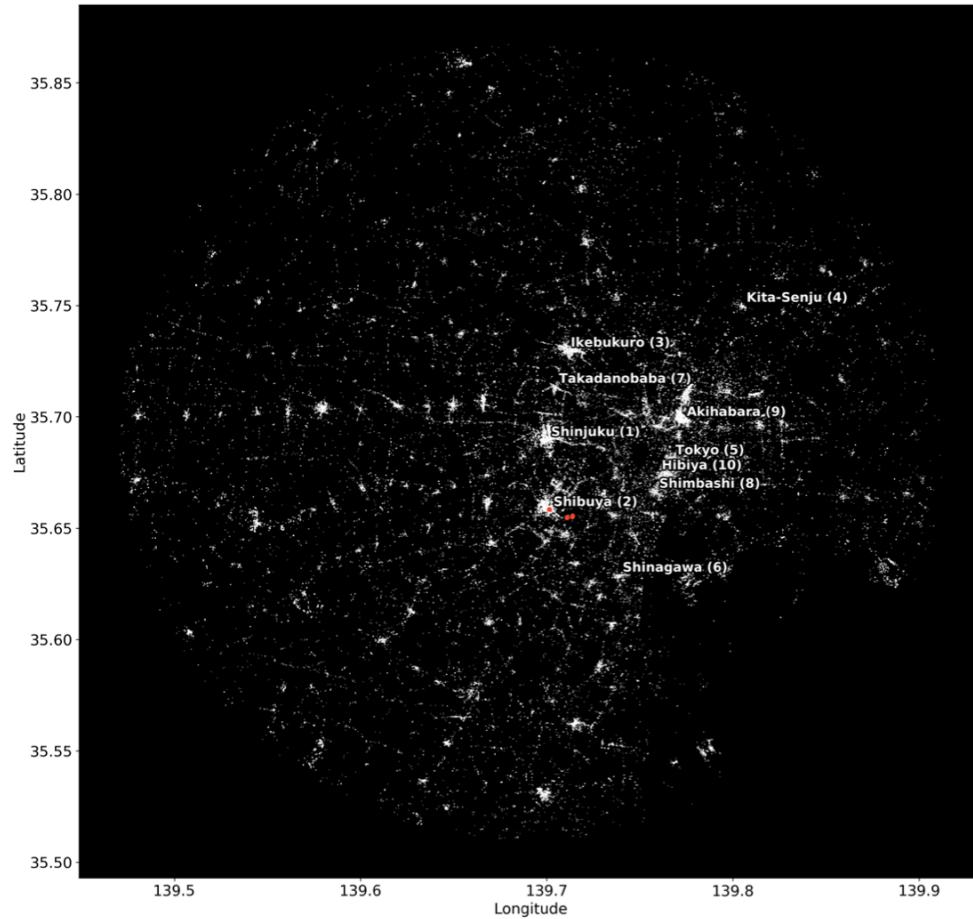
A Tale of Two Fridays: May 18 and 25

- Some venues are visited repeatedly by the same user.



A Tale of Two Saturdays: May 19 and 26

- Some venues are visited repeatedly by the same user.



Summary of Observations from EDA

- Check-ins concentrated near train stations and subway stations
- More check-ins during weekends
- Distinct check-in behavior between weekdays and weekends
- Imbalanced data set
- Check-ins concentrated near city center for weekdays, dispersed for weekends
- Individual users have somewhat regular temporal and spatial activity

Feature Engineering

- A user's past check-in history can be a good predictor of future check-ins due to temporal and spatial regularity.
- Then, for a user's current time and location, find the intersection of:
 - All check-in history for the same time +/- 1 hour
 - All check-in history for the same location, within a 1-km radius
 - All check-in history for the same day of the week, weekdays or weekends
- The result: 9 features, one for each venue category

Feature Engineering II

- The observations and features look like this:

Observations
(Check-ins)

timestamp1	userid1	(lat1, long1)
timestamp2	userid2	(lat2, long2)
timestamp3	userid3	(lat3, long3)
timestamp3	userid4	(lat4, long4)
timestamp4	userid5	(lat5, long5)

Features
(All past userid check-ins +/- 1 hour of current time, within a 1-km radius of (lat, long), for weekdays or weekends, for each of the 9 root venue categories, normalized to [0,1])

Target variable (0-8)

Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Label

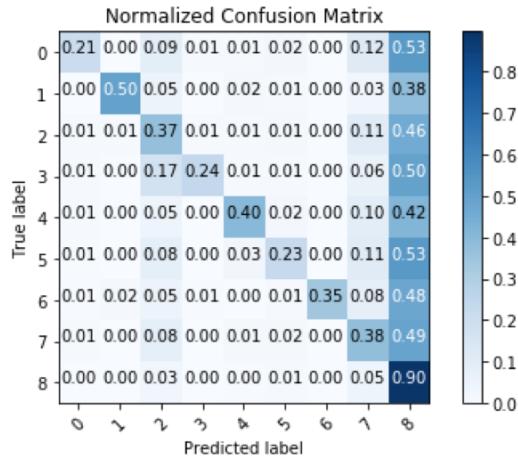
Supervised Learning Algorithms for Multiclass Classification

- Logistic Regression
 - Gaussian Naïve Bayes
 - K-Nearest Neighbors
 - Random Forests
 - Extreme Gradient Boosting
 - Extremely Randomized Trees
-
- 75:25 train-test split of 100,000 check-ins
 - Grid search to tune hyperparameters

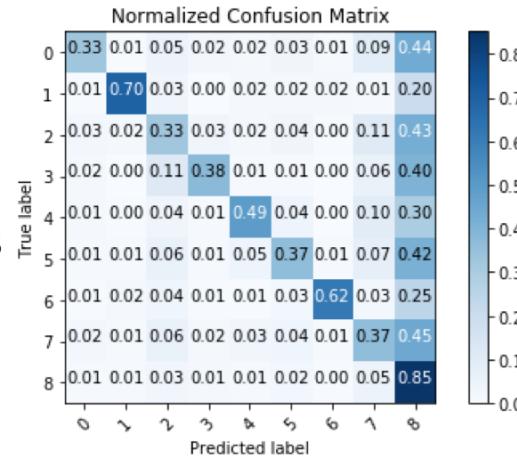
The confusion matrices of each classifier look similar...

- Accuracy: ~0.60. Many false positives for Label 8 (Travel & Transport)

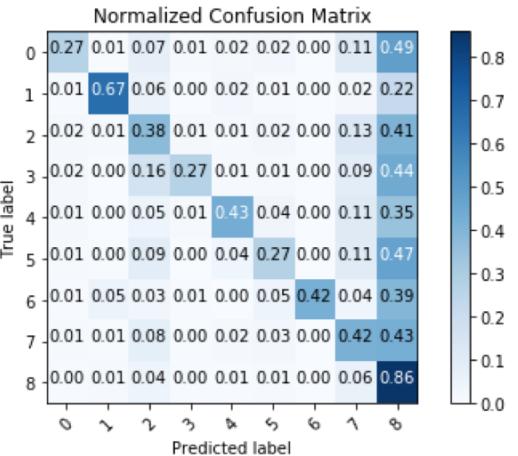
Logistic
Regression



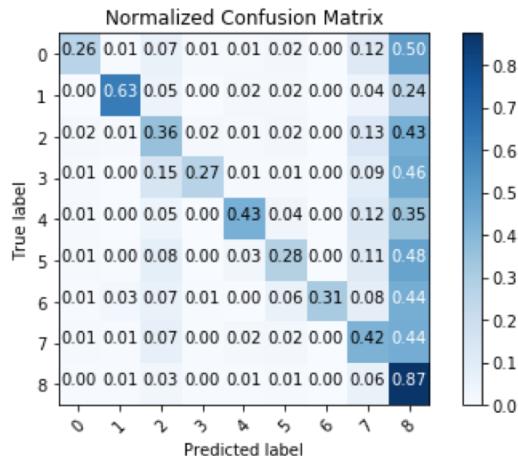
Gaussian
Naïve Bayes



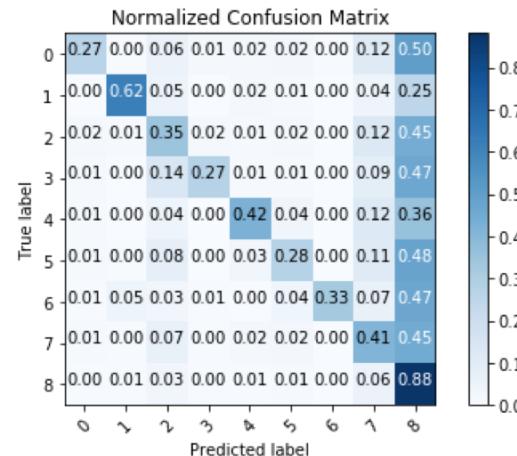
K-Nearest
Neighbors



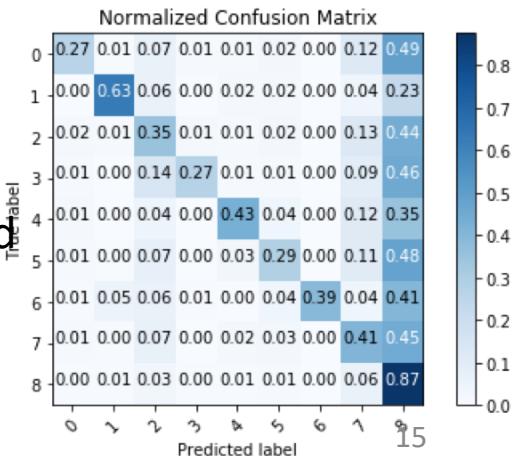
Random
Forests



Extreme
Gradient
Boosting



Extremely
Randomized
Trees

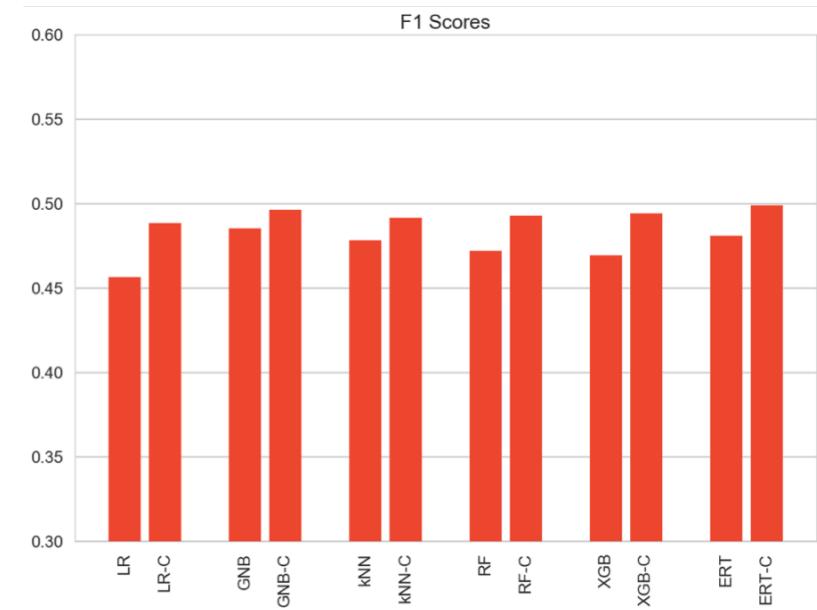
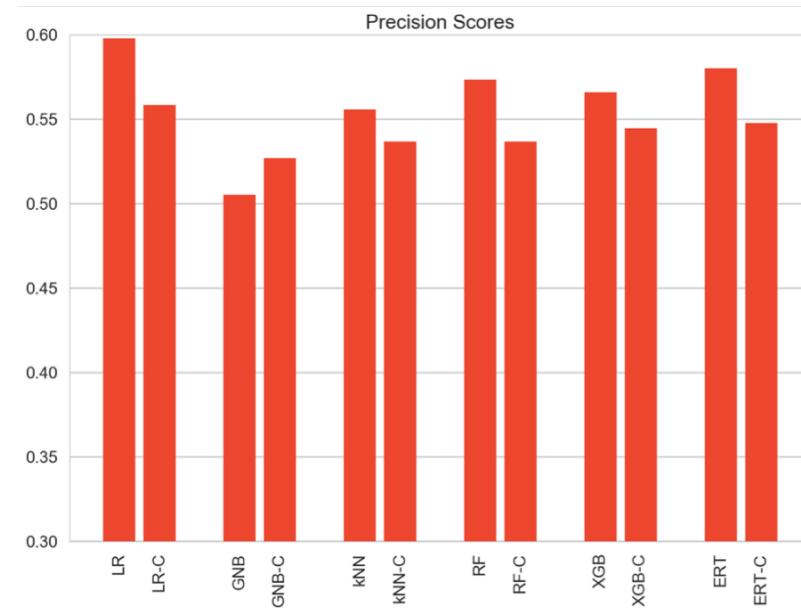


What can we do about the imbalanced data?

- Travel & Transport is the dominant venue category causing imbalance
- Can we perform binary classification first?
 - 1: Travel & Transport
 - 0: All other categories
- After classifier is trained and tested to yield predictions:
 - Remove all check-ins that are truly Travel & Transport from the train set
 - Remove all check-ins that are predicted Travel & Transport from the test set
 - With the remaining data, perform multiclass classification as before

Results of Cascaded Multiclass Classification

- Introducing cascaded classifiers improved all F_1 scores (+ 2%)!



Recommendations

- Introduce more features
 - Past history is not enough
 - What about factoring in order? What kinds of check-ins tend to appear consecutively for the user?
- Gather more data
 - More check-ins would prevent the situation that a user has no prior history for a particular time and location
- Introduce other data sources
 - Is weather a good predictor of check-in activity?

Conclusion

- Foursquare data in Tokyo was used to predict venue category of check-in based on previous check-in history for the same time and location.
 - Accuracy ~ 0.60
- Cascaded classifiers were employed to address the imbalanced data set.
 - F_1 score improved for every classifier by $\sim 2\%$
- Future work: Continue feature engineering, test model, and iterate...