Understanding the Behavioral Patterns of a Successful Driver

Data Science Institute
Columbia University

Authors: Ridhi Mahajan, Akshat Mittal, Nima Chitsazan, Foad Khoshouei, Ashwin Jayaraman
Industry Mentor: Zhiwei (Tony) Qin
Faculty Mentor: Vineet Goyal

Data Science Capstone Project with DiDi Labs

Project Introduction

The aim of the project is to identify the patterns which enable a driver to be successful on the DiDi platform. The ride hailing platform allows drivers the flexibility to choose when they start work and how long they stay online. However, it can be seen from the data that despite similar work hours and period of activity, some drivers are able to perform better than their peers. In this project, we seek to discover why.

Feature Extraction

To encode a driver's behavior, we extract spatial and temporal features. Temporal features help us evaluate whether rush hours have an impact on driver's performance. Spatial features help quantify the importance of pickup/drop-off location. We start with dividing Chengdu into square grids. Trying to better model the road-network, we move on to the use of hexagonal grids, and finally settle on polar coordinates to define concentric rings with equal ride-pickup densities.

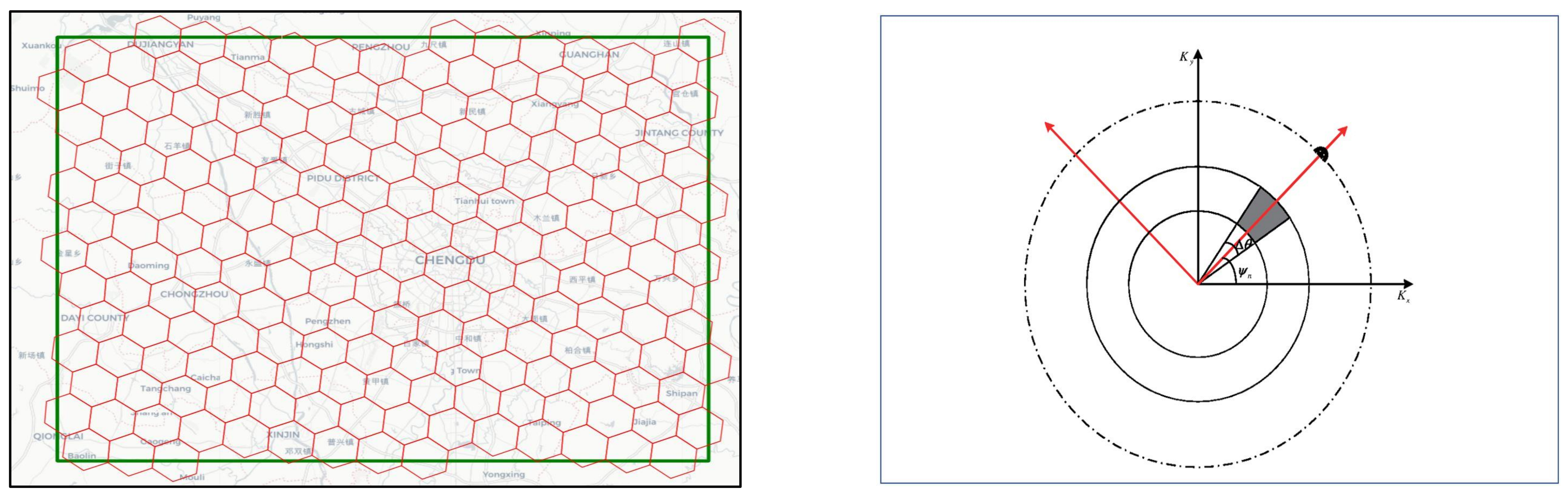

Figure 1. Spatial Features - Hexagonal Grids and Concentric rings

Driver Active Time Estimation

One of the most critical inputs to our model is the active time of a driver - the time that the driver is online on the app. We estimate this feature through various curve fitting and probabilistic approaches. We fit a gaussian through rides counts and then sample from an exponential distribution to approximate the 'patience' of the driver - how long a driver will wait for a ride before going offline.

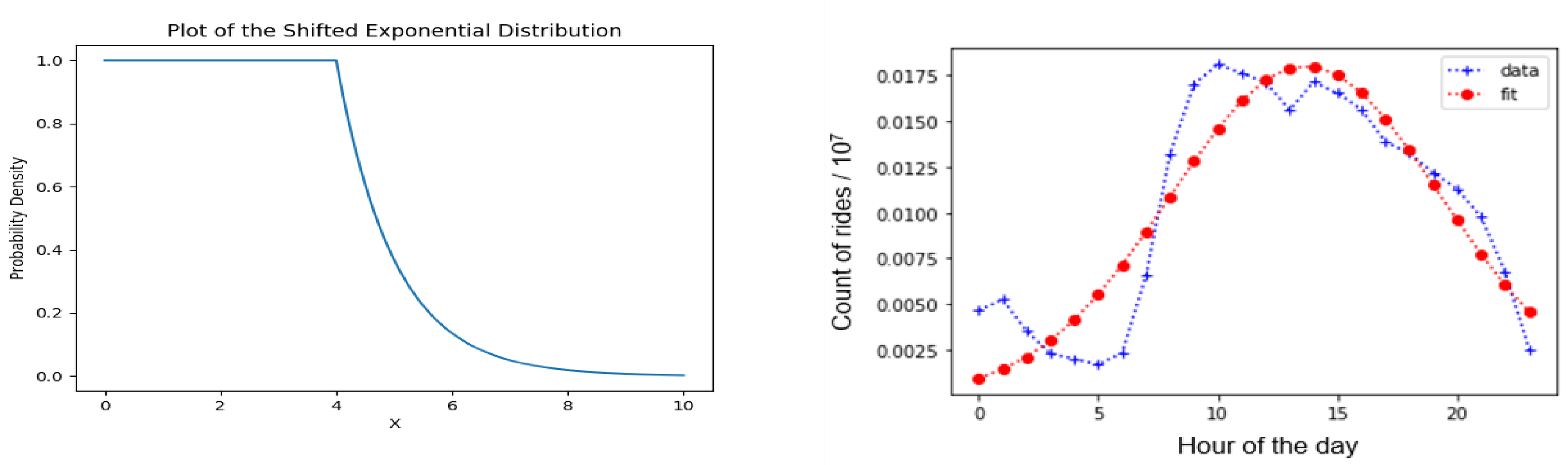

Figure 2. Distribution representing probability of being online as a function of time

Results

In the absence of direct earnings data, we use Performance Ratio - the ratio of a driver's ride time to active time - as the target. Random Forest Regressors are used to fit this target.

Index	Random Forest Model	Score
1	Active time estimated without threshold	0.48
2	Constant threshold	0.66
3	Time varying threshold (k = 1)	0.66
4	Time varying threshold (k = 10)	0.73
5	Time varying threshold (k = 100)	0.77
6	Time varying threshold and Radial Features	0.79

Table 1: Model Performance

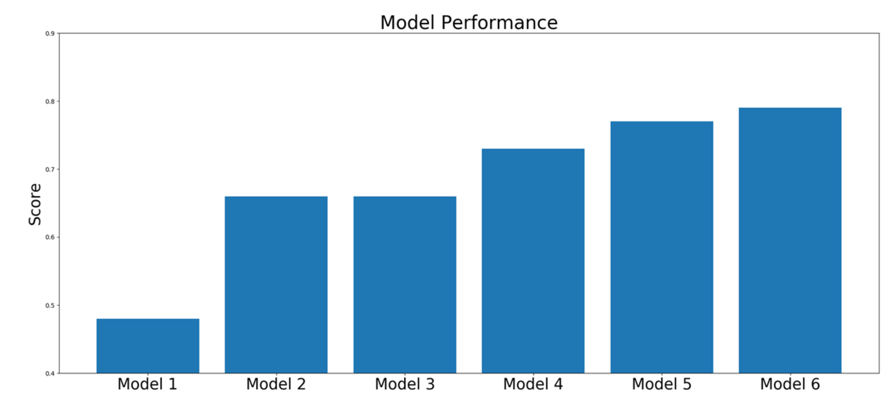

Figure 3. Model performance with different assumptions for active time

Feature Importance

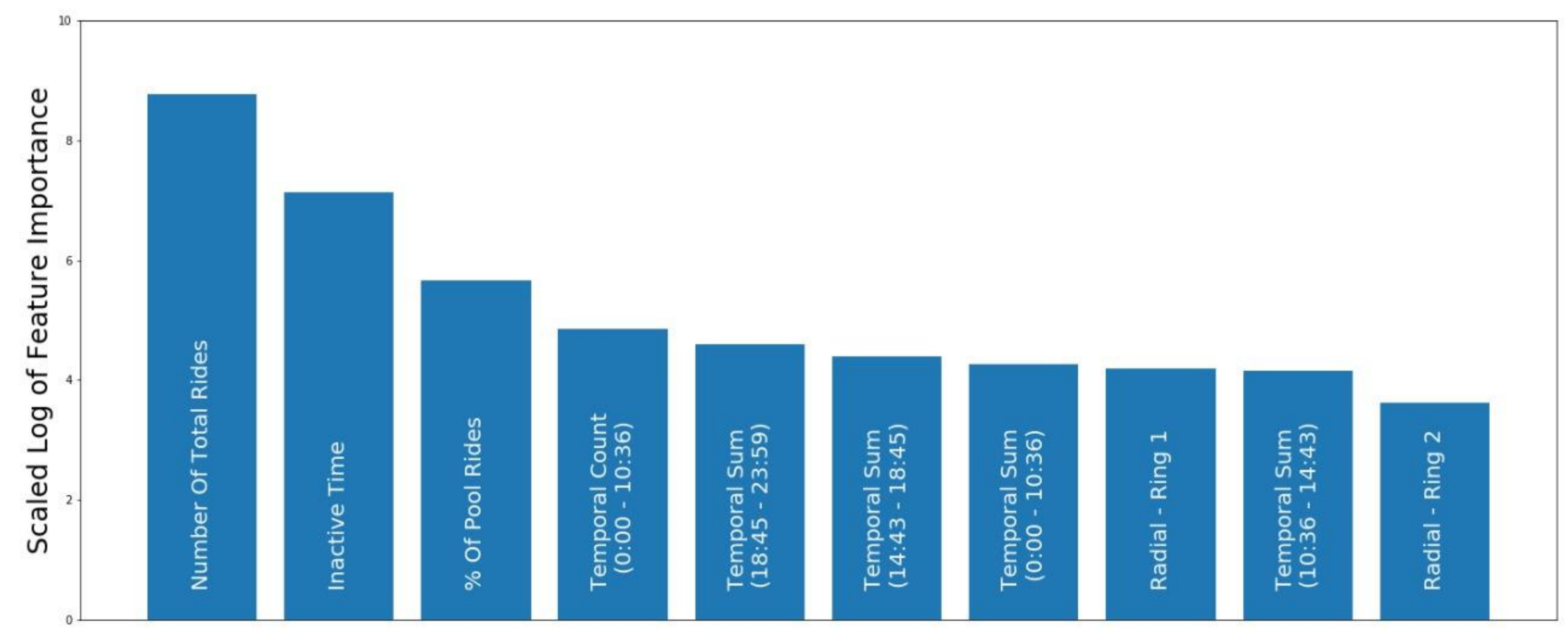

Figure 4. Scaled Log of Feature Importance derived from the Random Forest model

Conclusions

It can be seen from the above graphs that making the threshold a function of time improves the model significantly. Temporal features play an important role in determining driver performance. All else equal, strategizing when to be online can be crucial for being successful on DiDi's platform.

Acknowledgments

We are extremely thankful to our mentors, Professor Vineet Goyal (IEOR, Columbia University) and Zhiwei (Tony) Qin (DiDi Labs/ AI Labs) for their support and insight during the project.

References

Chen, L., Mislove, A., Wilson, C. Peeking Beneath the Hood of Uber. 2015. Proceedings of the 2015 ACM Conference on Internet Measurement Conference, pp. 495-508.

Liu, L., Andris, C., Ratti, C., Nov. 2010. Uncovering cab-drivers' behavior patterns from their digital traces. Computers, Environment and Urban Systems 34 (6), 541-548.