Uber app delivery time estimation

Maciej Szczutko Ewa Stebel

University of Wroclaw

January 29, 2024

Outline

Data description

- Dataset contains information about the provision of transportation services by Uber.
- Uber is a multinational transportation network company that operates a platform connecting riders with drivers through a mobile app.
- Data includes information on trips made in the USA in 2016 year.
- Data comes from Kaggle website.

START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
7/18/2016 10:37	7/18/2016 10:49	Personal	Cary	Morrisville	4.1	Moving
4/16/2016 15:10	4/16/2016 15:26	Business	Morrisville	Cary	6.1	Meal/Entertain
12/22/2016 23:27	12/22/2016 23:32	Business	Lahore	Lahore	2.1	Customer Visit
08-07-2016 17:28	08-07-2016 17:43	Business	Edgehill Farms	Whitebridge	2.7	Customer Visit
1/27/2016 14:05	1/27/2016 14:13	Business	Raleigh	Raleigh	2.7	Customer Visit

Figure: Head of UberDataset

Location visualisation

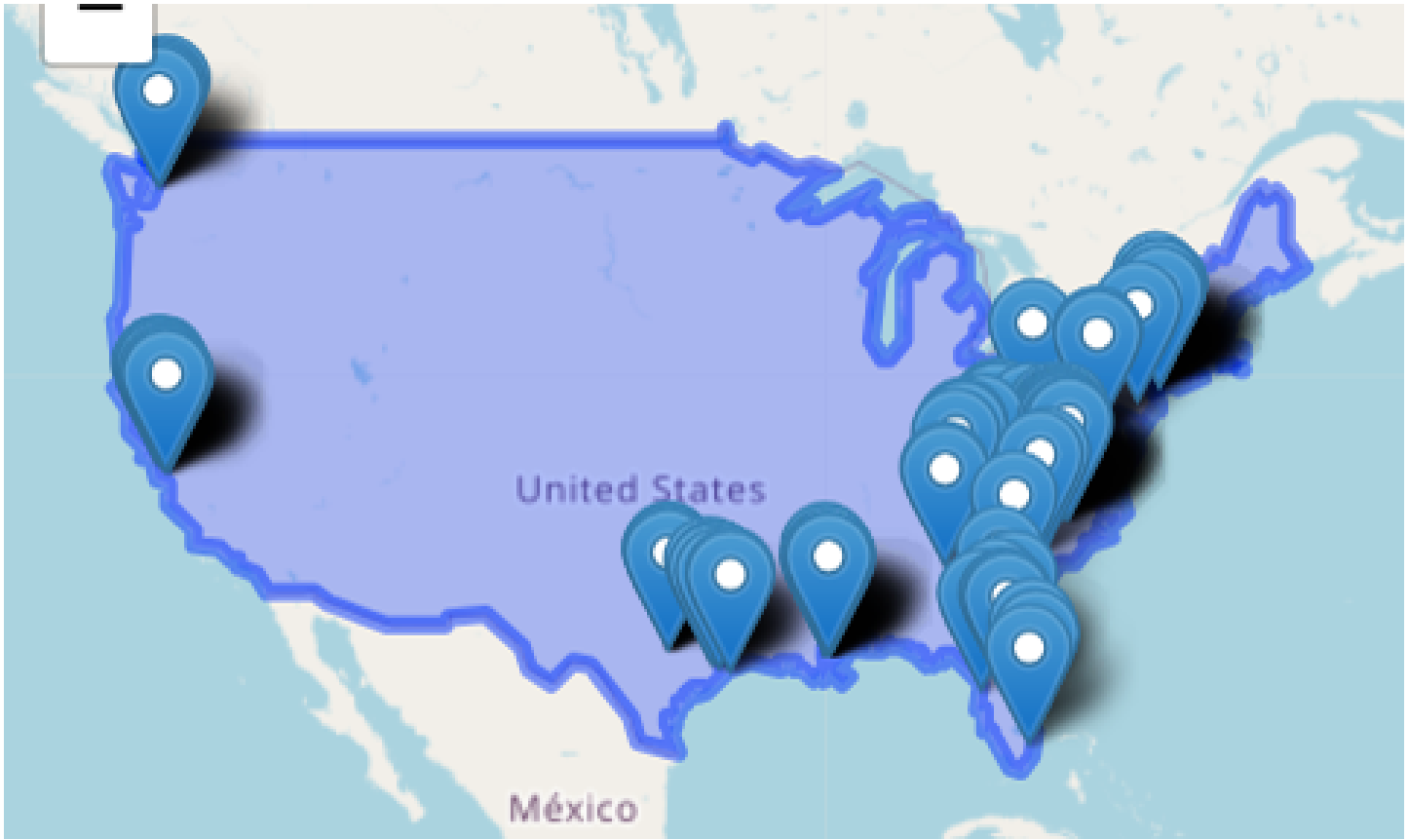

Figure: Location distribution

Bunch of stats

Statistic	Amount
Rows	1156
Columns	7
Categorical features	6
Continuous features	1

Encountered problems

- Missing data in some columns
- Inconsistent data types (but all valid, no need to removal)
- Garbage data (cancelled trips, 0 duration, etc.)
- Copyrights :)

Histogram of locations

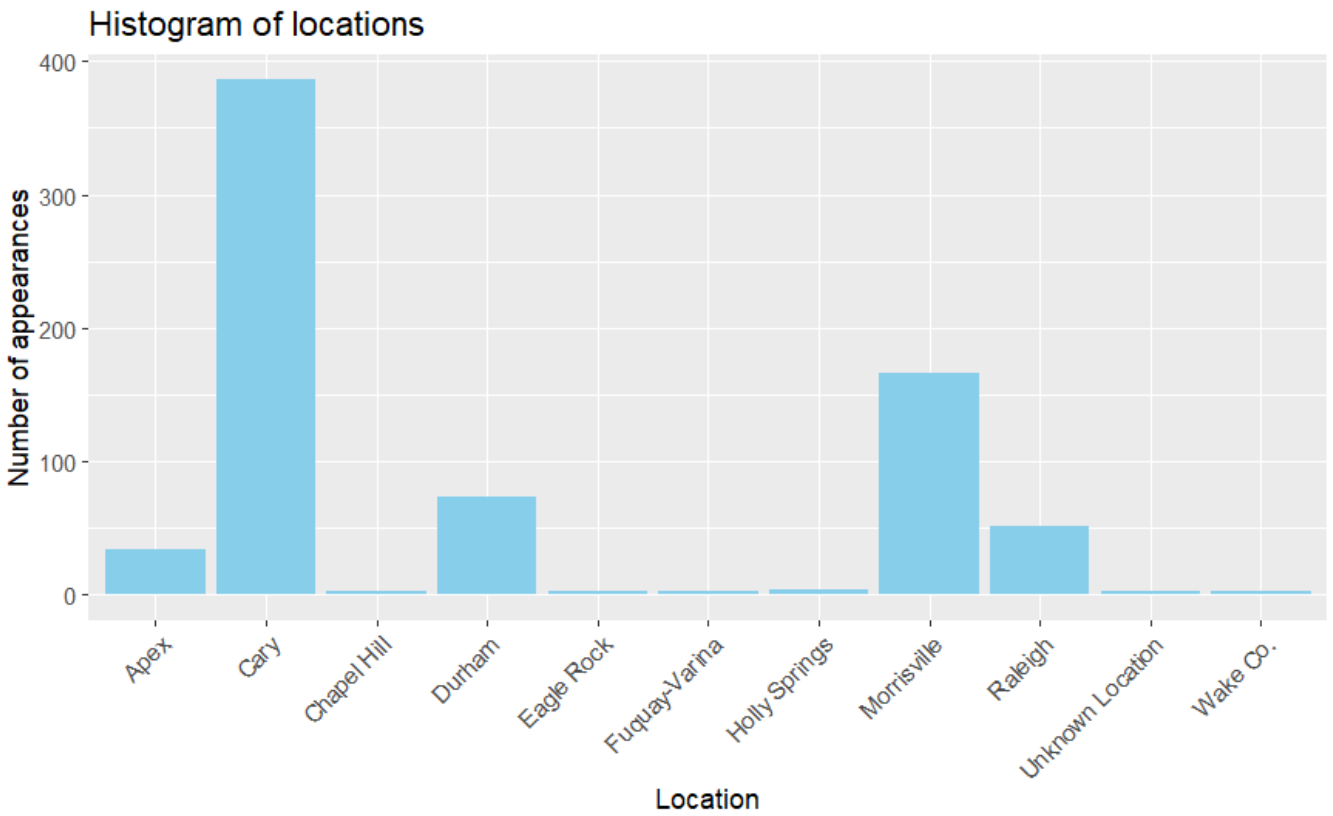

Figure: Histogram of locations

Data distribution

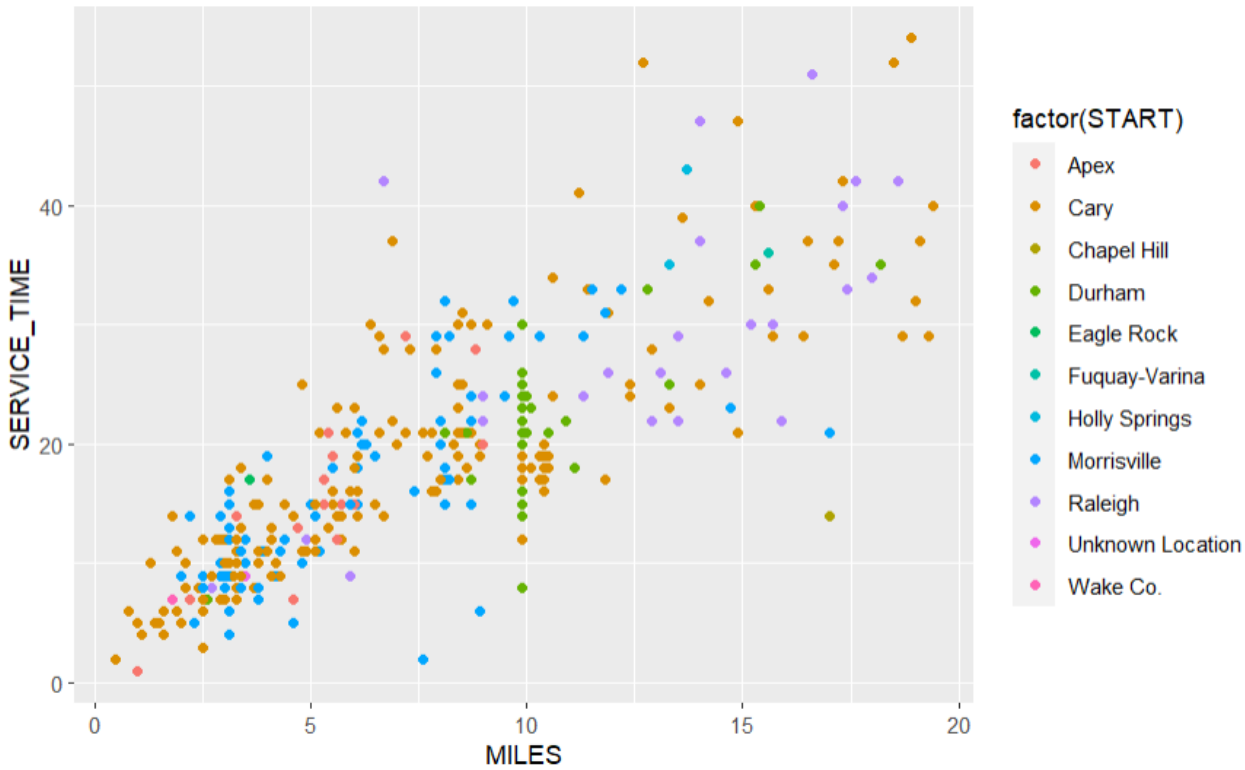

Figure: Chart of service time by miles and location

Dependence on the hour and time of day

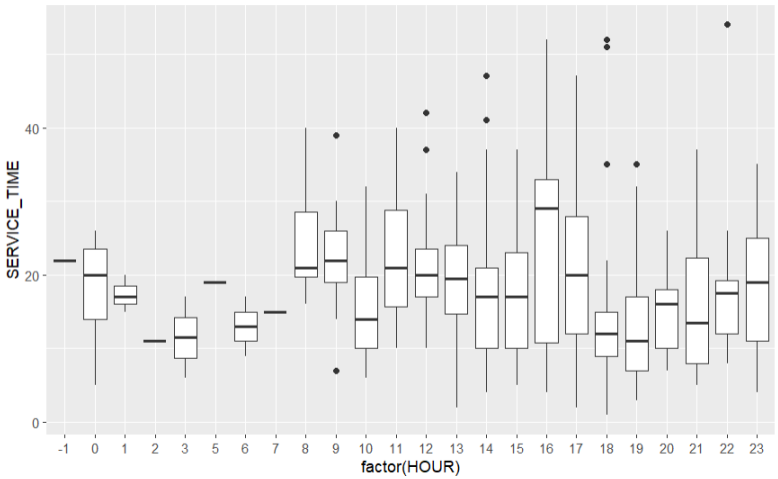

Figure: Boxplot of hour variable

Figure: Chart of daytime variable

Dependence on the purpose and category

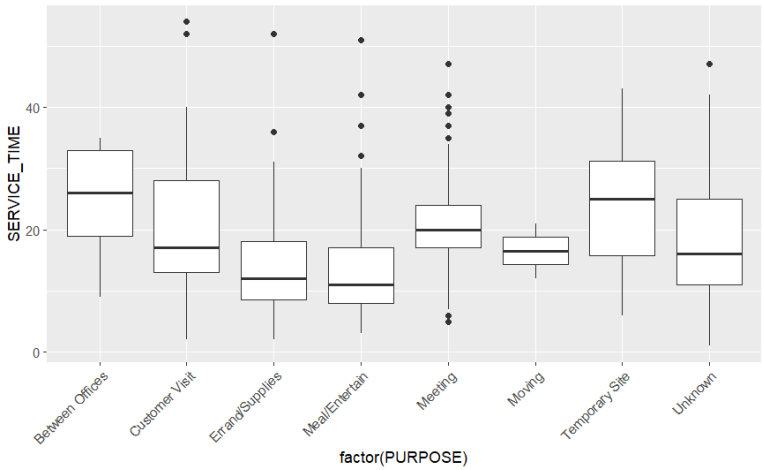

Figure: Boxplot of purpose variable

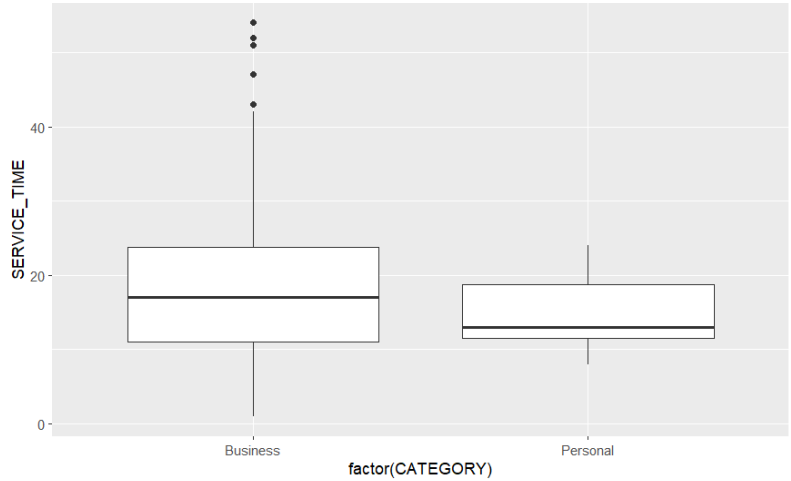

Figure: Boxplot of category variable

Feature engineering

We want to use information about what the time of the day is (e.g. rush hours, night etc.) and use it as the extra input for the models. To do this we want to use just hours. But 24 levels seems to be quite big number considering volume of data and fact some of them truly do not make any difference. There is no ready to use framework to bucketing categorical variable (in regression problem) so we try to based on intution on boxplot.
We test a lot of possibilities but finally, we end up with new variable DAY_TIME that indicate wheter we have rush hour (15-18) or not.

Used Models

The idea

Based on above plots and distribution we want to check two approach. First one – where we manually select predictors for model (rather simple one) and second one – where we use framework to select best model. We used GAM and ordinary LM models.

- LM : $SERVICE\_TIME \sim MILES$
- LM :$SERVICE\_TIME \sim MILES * DAY\_TIME$
- GAM :$SERVICE\_TIME \sim s(MILES)$
- GAM :$SERVICE\_TIME \sim s(MILES * DAY\_TIME)$

Example models

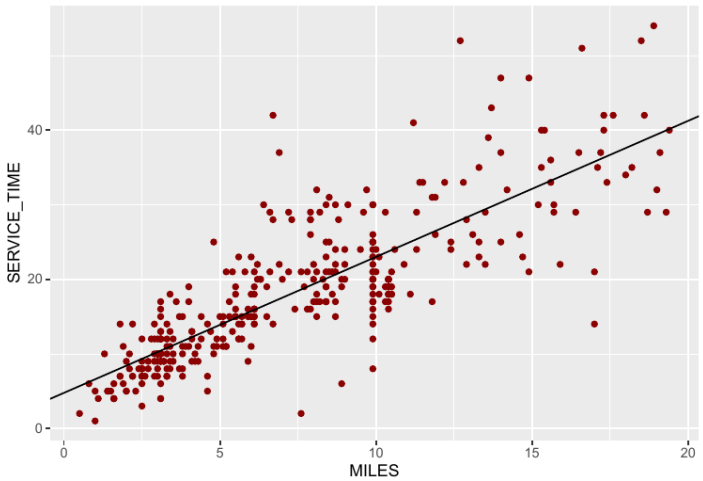

Figure: Ordinary LM model

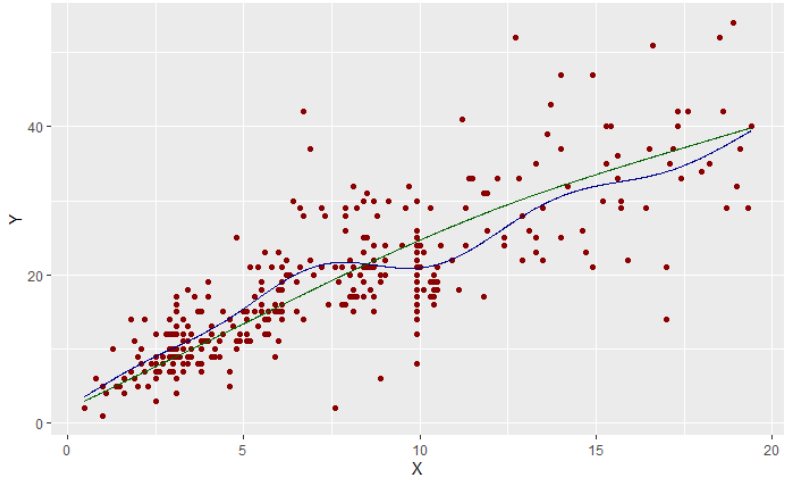

Figure: Interaction between *MILES* and *DAY_TIME*

Metrics used

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Models performance

Result based on CV with k=5 folds.

Table: Mean Values for RMSE, MAE, and R^2 for Each Model

Model	Mean RMSE	Mean MAE	Mean R^2
LM	5.64	4.05	0.673
LM with interaction	5.59	3.98	0.683
GAM	5.53	3.93	0.702
GAM with interaction	5.61	3.98	0.707

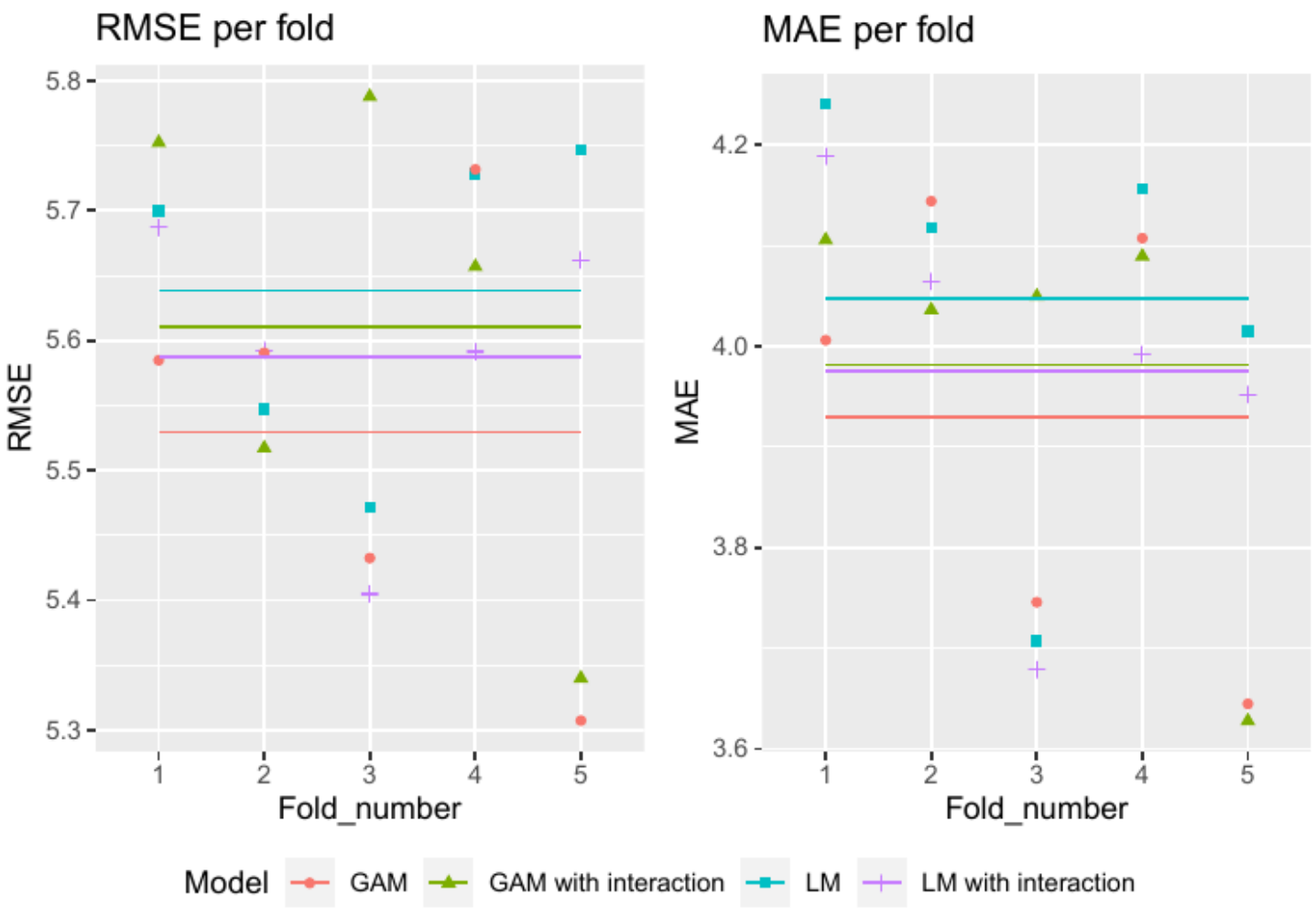

Figure: Metrics plots

Thank You

Thank you for your attention!