**PROBLEM STATEMENT**

The RIDE is MBTA’s door-to-door shared-ride paratransit service managing large complex operations.

- **$115M** Annual budget in 2019
- **1.65M** Number of trips in 2019
- **1238** Number of drivers trained by 4 providers

**Objectives:** Quantify and include driver behavior in the RIDE’s operations management schemes and reduce demand and supply mismatch.

**Problem definition**

February-March

**Feature analysis and engineering**

April-May

**Driver score design and delivery**

June

**Demand forecasting model**

July

---

**DRIVER BEHAVIOR**

**Quantifying behaviors**

Our exploration analysis on the GPS data lead to three comprehensive categories of driver behaviors:

1. *Drivers’ Deviations from The RIDE’s schedule*
2. *Drivers’ Breaks and start behaviors*
3. *Drivers’ Interactions with the system*

**Driver score definition**

Two goals were defined for a driver score: Capture the most important behaviors and differentiate drivers based on their performance. A survey conducted with the RIDE’s managers assessed the importance of each feature. Hence, for driver j on day d we get the following score:

\[
\text{score}_{j,d} = -1 \times \sum_{i=1}^{n} \text{Behavior}_{i,j}(\alpha_i + \beta_i) / \sum_{i=1}^{n} (\alpha_i + \beta_i)
\]

Higher score for better drivers

**Case study**

Driver score catches the overall behavior and enable comparison while each provider has its own distinctive behavioral patterns.

---

**DATA SOURCES**

Transportation and scheduling data sources requiring massive data computing:

- **90M** GPS points every 2 minutes indicating driver’s position, speed and system interaction
- **2M** Trips data on scheduled trips, their timings and associated driver
- **200K** Routes data on the overall system schedule

Quarterly run-structure: supplied driving hours in the overall network

---

**DEMAND FORECAST**

**Identifying non-revenue time**

In the overall network, we identified that 50% of the supply time is spent without any client interaction. The supply is defined by hour, by provider, by route lead to the idea of a geographic approach:

**Geographic clustering**

K-means clustering helped us identify 8 stable geographical clusters for trip pick-ups and departures.

**Model building**

Timeseries’ analysis with *tsfresh* and gradient boosting model for prediction:

- Training on 2018: June 1st to Nov. 23rd
- Testing on 2019: June 1st to Nov. 23rd

**Results**

Maintaining high-resolution prediction with only one month of prior observations:

- Prior data period: Six months, One month
  - Operating time: 70.1%, 69.1%
  - Number of trips: 79.5%, 78.8%

---

**BUSINESS IMPACT**

- Link with the garage location for drivers depending on geographical demand.
- New design of the run-structure precisely identifying which areas are served at each time-bin.
- Moving from a scheduling system based on optimizing only to a system based on prediction-prescription methods.

---

**Links**

- [John Nicholas & El Ghali Zerhouni](#)
- [Faculty Advisor: Prof. Dimitris Bertsimas](#)
- [The RIDE advisors: Abhishek Rai & Christopher Jurek](#)