

Machine Learning in Real-Time

Predicting taxi fares in NYC with Dataiku

Alex COMBESSIE

March 2019

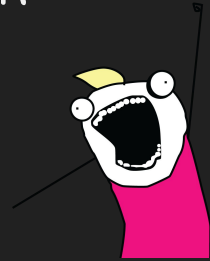


Intro: Why Real-Time is the New Black

REAL-TIME ANSWERS: GOOGLE

REAL-TIME TRANSPORT: UBER

REAL-TIME MACHINE LEARNING:



Agenda

1. In Search of Good Features
 - 2. Traaaaaaining Time!**
 3. Exposing our Model to Users... In a Real-Time App
- Outro: Lessons Learned & a Few More Things



In Search of Good Features



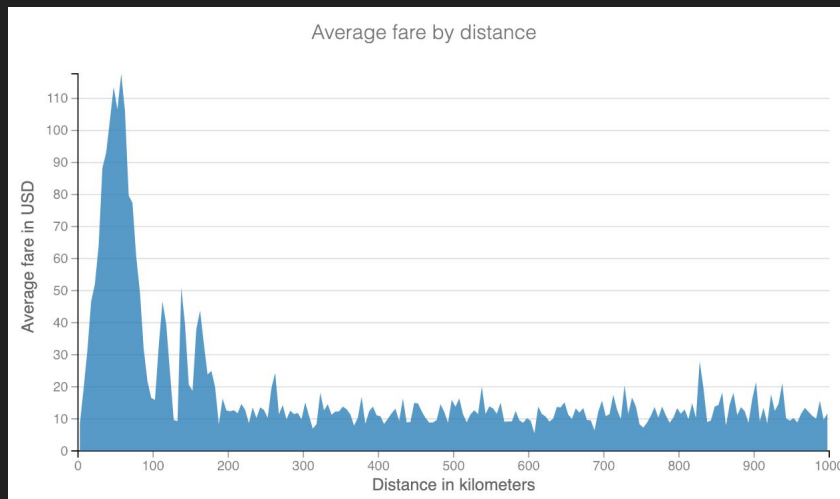
Discovering the Data at Hand

Poor, Dirty Data

www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

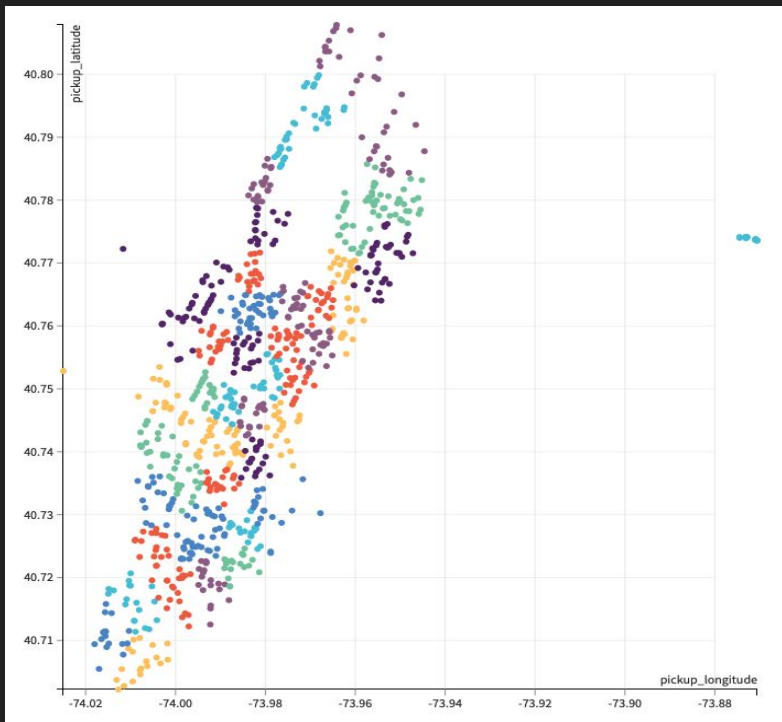
- **Only 4 raw features:** pickup time and location, drop-off location, number of passengers
- **Weird stuff:**
 - 1.9M rides < 100 meters?
 - 100K rides > 300 km??
 - To the bottom of the Hudson???

Non Linear Relationships



Making Features by the Hundred

What Do You See?



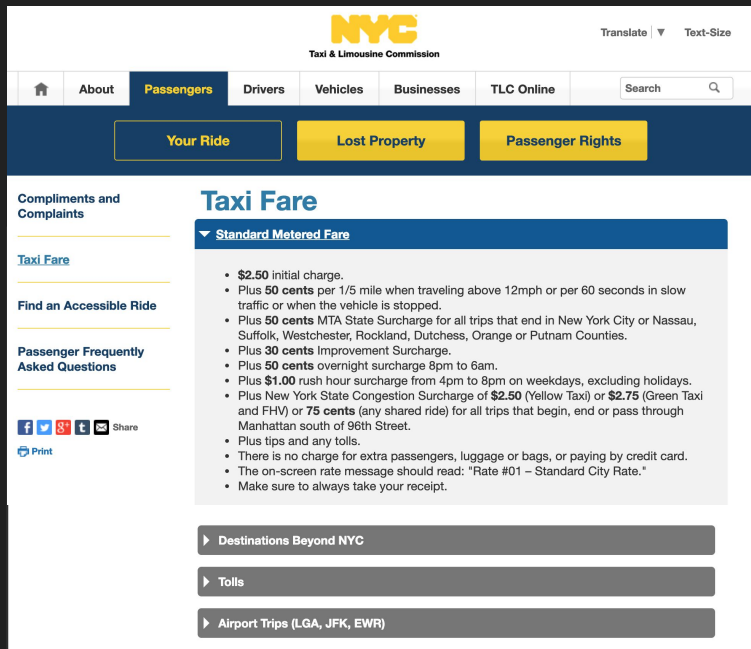
Feature Engineering x Many Iterations

1. **Start simple:** geometrical distances, time seasonalities
2. **Build up with unsupervised learning:** Clustering on GPS coordinates to assign pickup/drop-off to neighborhoods
3. **Finish with windowing:** aggregate features like avg/max fare from one cluster to another in the last 10/100/... rides




Going Back to the Root Cause

The Not-So-Secret Formula



The screenshot shows the NYC Taxi & Limousine Commission website. The header includes the NYC logo, a 'Translate' dropdown, and a 'Text-Size' link. The navigation bar has links for 'Home', 'About', 'Passengers' (highlighted), 'Drivers', 'Vehicles', 'Businesses', and 'TLC Online'. Below the navigation bar are three buttons: 'Your Ride', 'Lost Property', and 'Passenger Rights'. The main content area is titled 'Taxi Fare' and features a 'Standard Metered Fare' section with a list of charges: \$2.50 initial charge, 50 cents per 1/5 mile, 50 cents MTA State Surcharge, 30 cents Improvement Surcharge, 50 cents overnight surcharge, \$1.00 rush hour surcharge, New York State Congestion Surcharge, tips and tolls, and a note about extra passengers. Below the fare list are three expandable sections: 'Destinations Beyond NYC', 'Tolls', and 'Airport Trips (LGA, JFK, EWR)'.

Last Round of Enrichment

- Convert the formula into features
 - Flags for specific areas and hours
 - Tolls & airport trips
 - Traffic conditions → the  API

Focus on What Matters

- Too many features: 4 ↗ 500 ↘ 100 (highest correlated features)





Traaaaaining Time!



Choose Your Weapon Algorithm

The Arsenal

Random Forest	<input checked="" type="checkbox"/>
Gradient tree boosting	<input checked="" type="checkbox"/>
Ordinary Least Squares	
Ridge Regression	
Lasso Regression	
XGBoost	
Decision Tree	
	Support Vector Machine <input checked="" type="checkbox"/>
	Stochastic Gradient Descent <input checked="" type="checkbox"/>
	KNN <input checked="" type="checkbox"/>
	Extra Random Trees <input checked="" type="checkbox"/>
	Neural Network <input checked="" type="checkbox"/>
	Lasso Path <input checked="" type="checkbox"/>
+ ADD CUSTOM PYTHON MODEL	



Microsoft
LightGBM

- **Better = - 0.3 RMSE**
- **Faster = x 3 speed**
- **Stronger = not Out of Memory**



Fighting the Evil Dr. Overfitting

By Feature

- **Balanced view of all features VS always learning from the most predictive**
- **Reduce “colsample_bytree” parameter to a lower percentage of 60% instead of 100% – a.k.a. Bagging**
- **Grid-Search**

By Observation

- **Generalize to the entire dataset VS specific to a small group of observations**
- **Limit tree growth by setting a “min_split_gain” threshold in addition to “max_depth”**
- **MOAR Grid-Search**



Exposing our Model to Users... In a Real-Time App

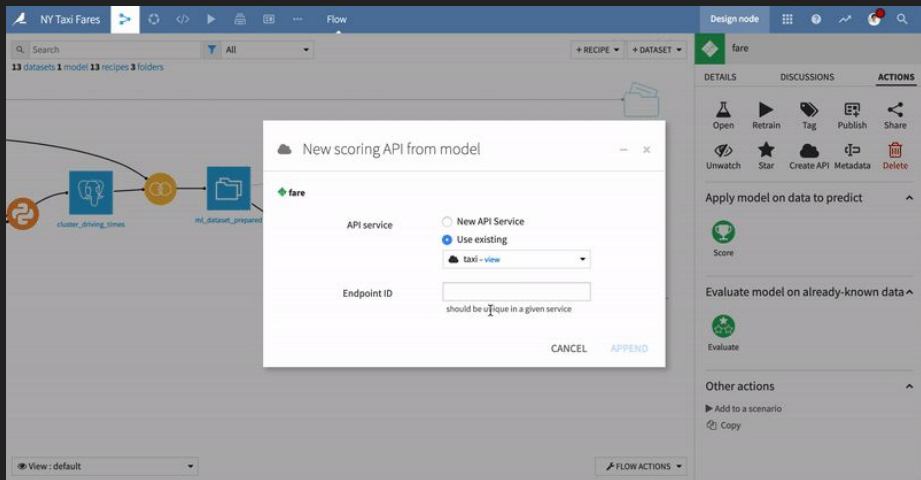


From Batch to Real-Time (API)

API Service Structure

`predict_fare`: Python endpoint to take raw features and output the fare prediction, a "wrapper" to call...

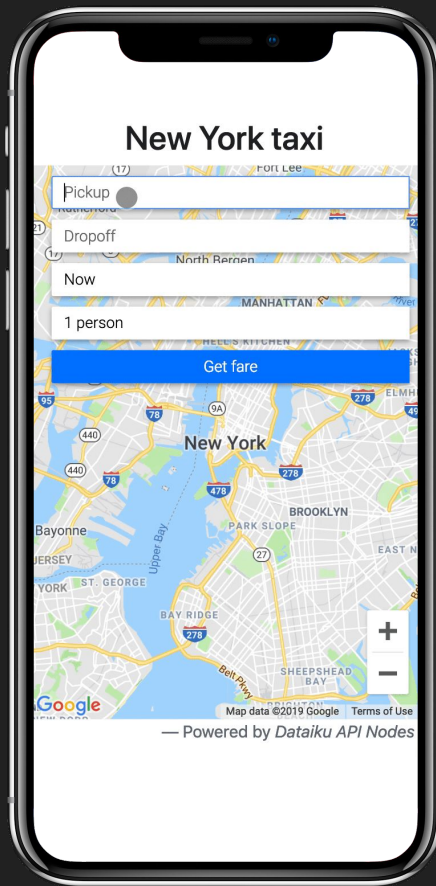
- A. `_cluster`: Python endpoint to assign pickup and dropoff to neighbourhood clusters
- B. `_traffic`: Python endpoint to get traffic data from the HERE API
- C. `_fare`: our predictive model using both raw features and traffic data



Turning a model into an API,
as easy as pie!



Just Demo



taxifare.dss-demo.dataiku.com



Outro



4 Things You Can Learn by Doing

1. Understand the problem before building models
2. Do not add features for the sake of features
3. Try as many algorithms as possible
4. Simplify your pipeline before deployment



データ 育

*From English data
and Japanese affix -iku (育)
"To raise or bring up; to grow up"*

*Literally, "Data Education"
or "Let's Grow the Data skills"*



Join Us at EGG LDN

The Human-Centered AI Conference

July 2, 2019

london.egg.dataiku.com

Early Bird Discount: MancML



Thanks!

Questions?

Try it yourself

dataiku.com/dss/trynow



Google
Cloud



Microsoft Azure

... or any Linux
server/container

Learn and stay in touch

academy.dataiku.com

mailto:

alex.combessie@dataiku.com

follow: [@alex_combessie](https://twitter.com/alex_combessie)

