SF BikeShare

Akanksha, Marine, Esther, Lexie

Dataset

SF Bikeshare Data : 2GB

Station

Status

Trips

Weather

SF Population Data

Population by Neighborhood in San Francisco

Analytic goals

Analytic goals

Predict number of bikes available at a given station with

- station information
- weather condition
- type of day
- hour
- population

Related Works

- Predicting number of daily trips
  - Predictors:
    - weather condition
    - number of bikes available
    - type of day (business day vs holiday vs weekend)
- Most important features:
  - business_day, temperature, month

Source: https://www.kaggle.com/currie32/a-model-to-predict-number-of-daily-trips

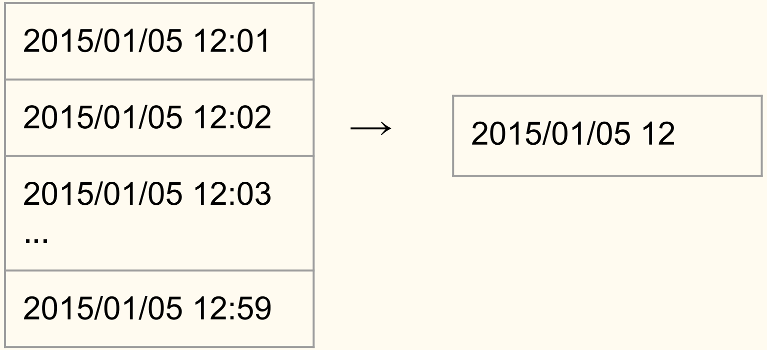
Preprocessing Algorithms

Total runtime: **31 mins**

Weather:

- *precipitation* → *is_rain*
- *mean_temperature*

Status:

- Convert *date* to DateTime() → *is_weekend*
- 

  2015/01/05 12:01

  2015/01/05 12:02

  2015/01/05 12:03
  ...

  2015/01/05 12:59

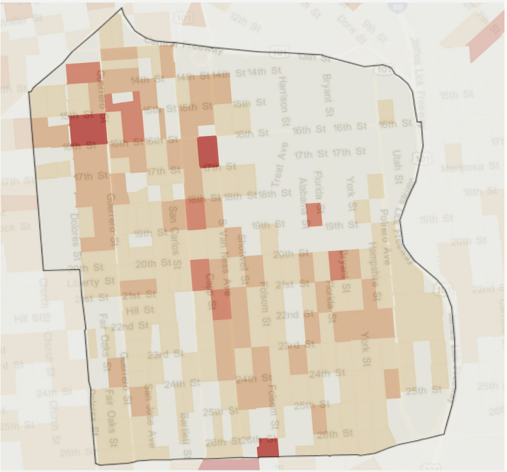
  →

  2015/01/05 12

Preprocessing Algorithms cont...

Adding **population** field:

- Station table has lat, long columns
- Used Socrata SF Data API & GeoNames API to map the lat, long columns to SF neighbourhoods
- Joined neighbourhood population data to the the station table



Machine Learning Outcomes

Spark ML

Linear Regression: 3.4037

Decision Tree: 2.92

Gradient-Boosting: 2.93

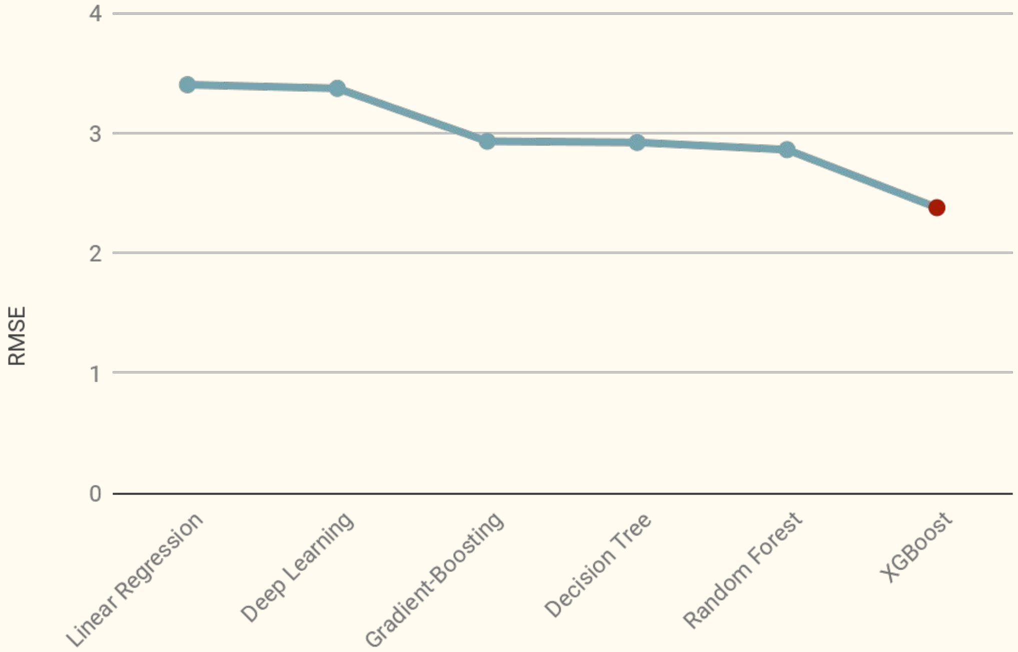
Random Forest: 2.86

H2O

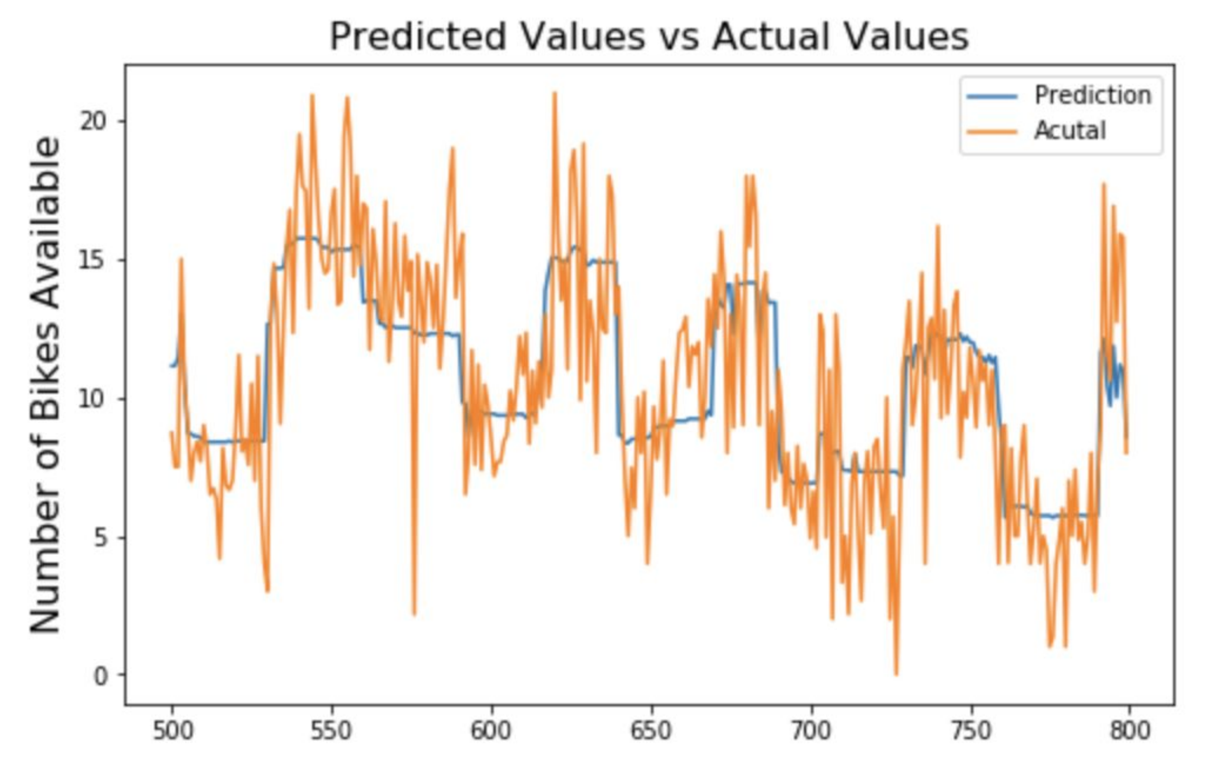
Deep Learning: 3.37

AutoML - XGBoost: 2.71

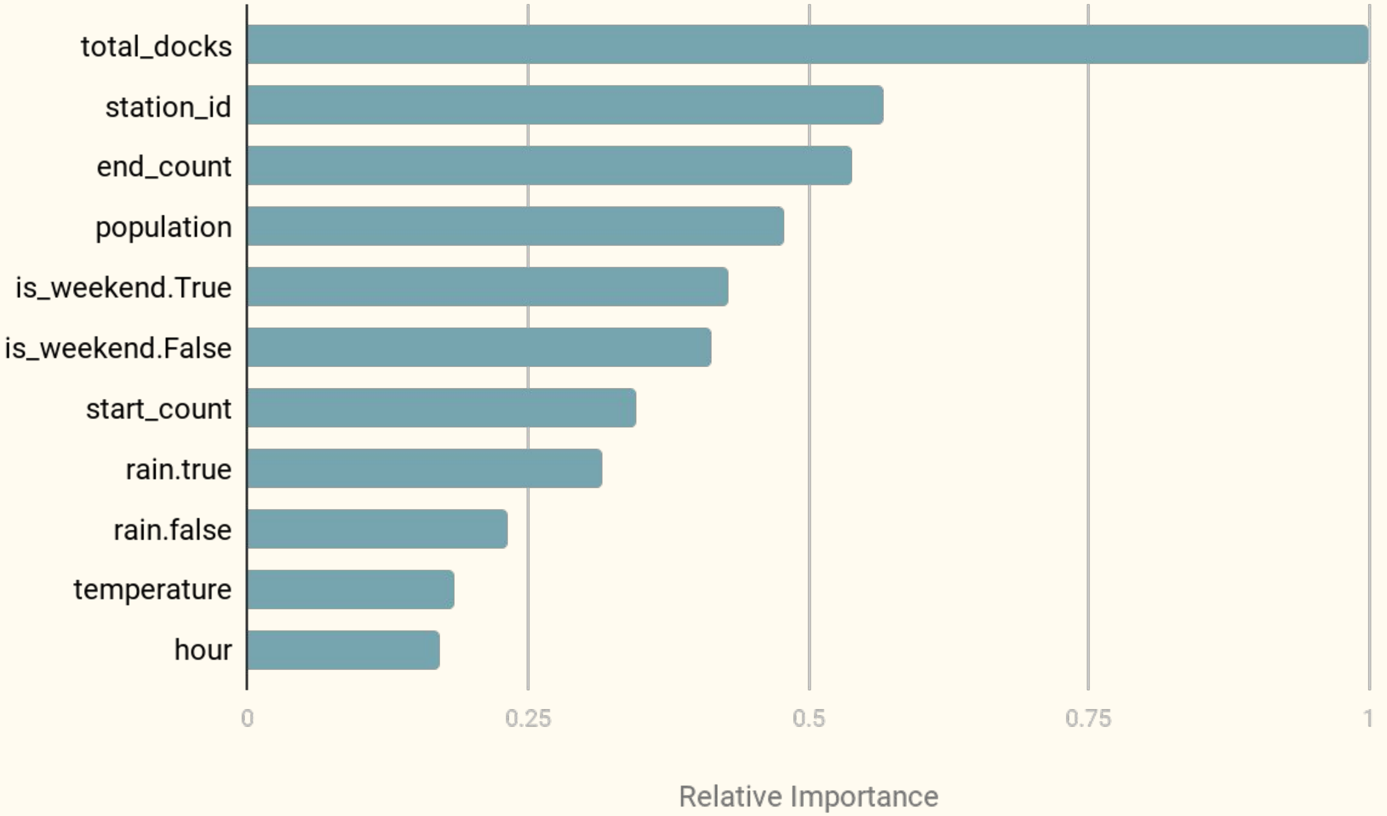
XGBoost Regressor has the lowest RMSE



Random Forest Predictions



Feature Importance



H2O

Deep Learning:	Train RMSE	Test RMSE
	3.48	**3.37**
Auto ML - XGBoost:	Train RMSE	Test RMSE
	2.22	**2.37**

Runtime Comparison

	r5a.8xlarge (memory optimized)	**r5a.12xlarge** (memory optimized)	**c4.8xlarge** (compute optimized)	**c4.8xlarge** (compute optimized)
Nodes	5 nodes	3 nodes	4 nodes	5 nodes
Time	14m 12s	13m 2s	12m 15s	11m 43s

Conclusion & Lessons Learned

- For spark ML: Random Forest performed the best on our data
- For H2O: XGboosting performed the best
- H2O was slower on EMR clusters when operating on data
- Population of the station area is correlated to the number of bikes available