Predict New York City Taxi Demand

Capstone Project Big 4 Team 17/09/2016

Big 4 Team

Shuo Zhang

PHD Chemical Engineering

Bin Fang

PHD GIS

Jingyu Zhang

PHD Electrical Engineering

Yunrou Gong

Master Operational Research

Content

- Introduction
- Data Pre-processing
- Modeling Workflow
- Exploratory Data Analysis
- Modelling
- Ensemble
- Prediction
- Future Work

Introduction

Purpose: predict the number of taxi pickups given a one-hour time window and a location within NYC

- how to position cabs where they are most needed
- how many taxi to dispatch
- how ridership varies over time

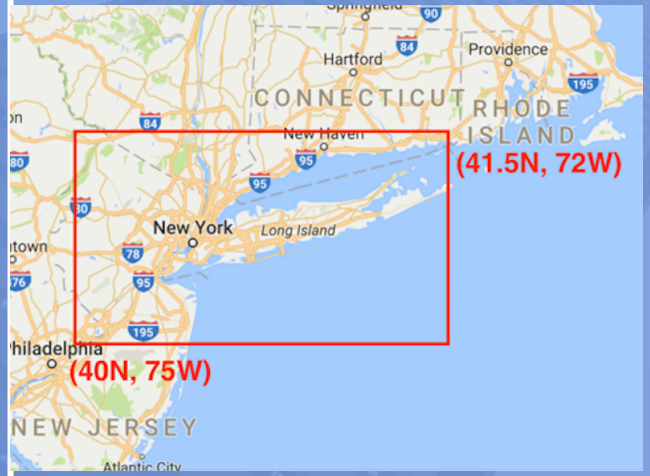
Input:

Date, one-hour time window, and zipcode within NYC

Output:

Predicted number of taxi pickups at the input time and location

Data Preparation



Green 1.2G

Uber 0.22 G

Yellow 12.5G

Weather 0.001G

Raw Data ~14G

Data Cleaning

Feature selection ~1billion

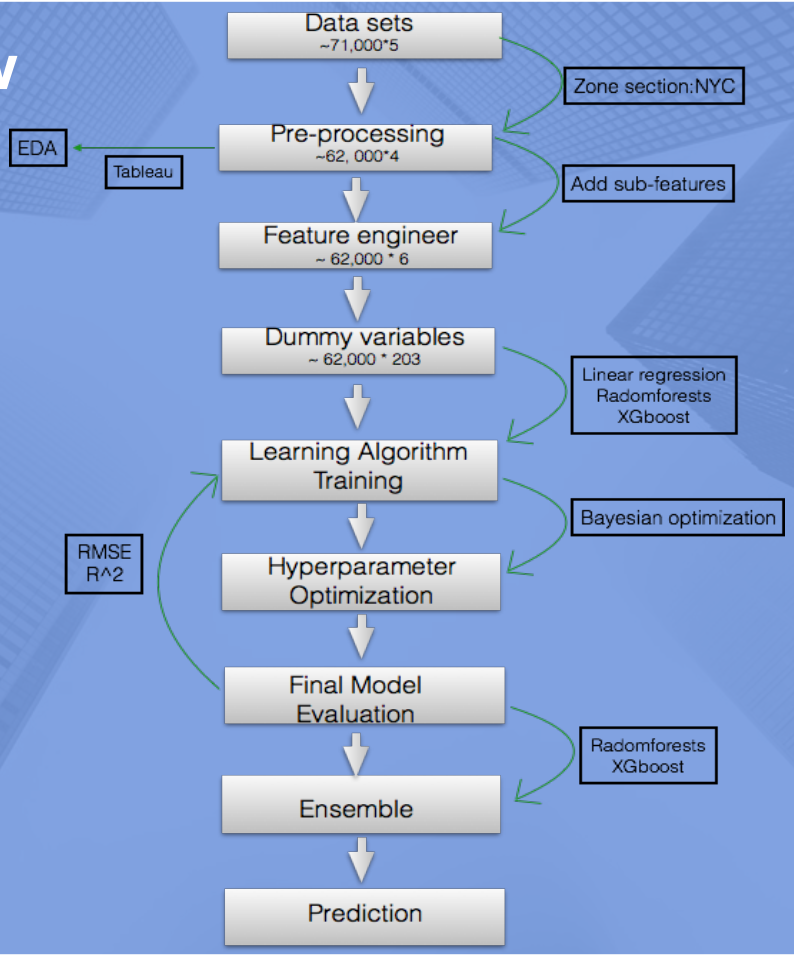
Lon,lat-->zipcode Geocoding 120,000

Dimension reduction ~1 billion

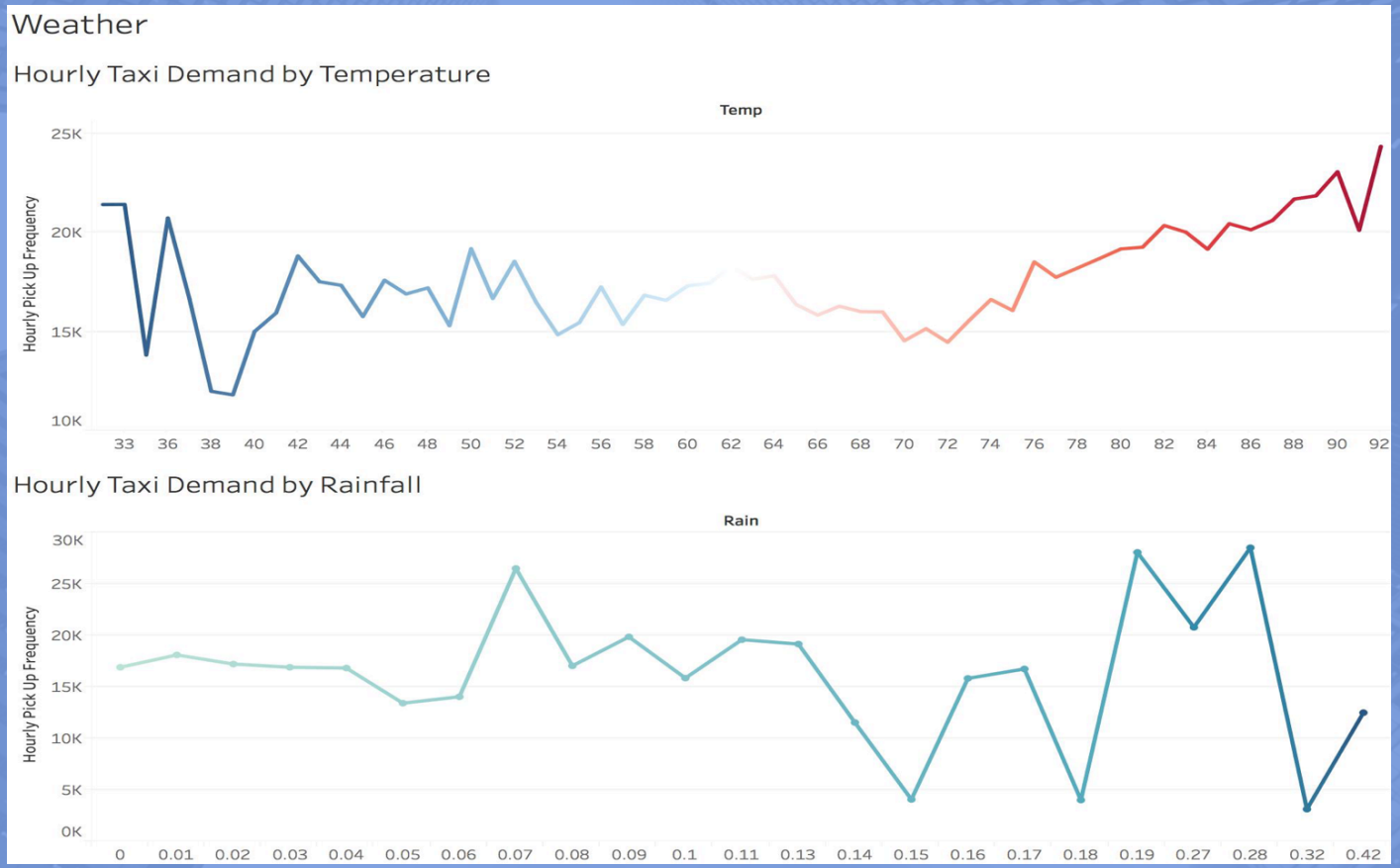
Aggregation ~ 60,000

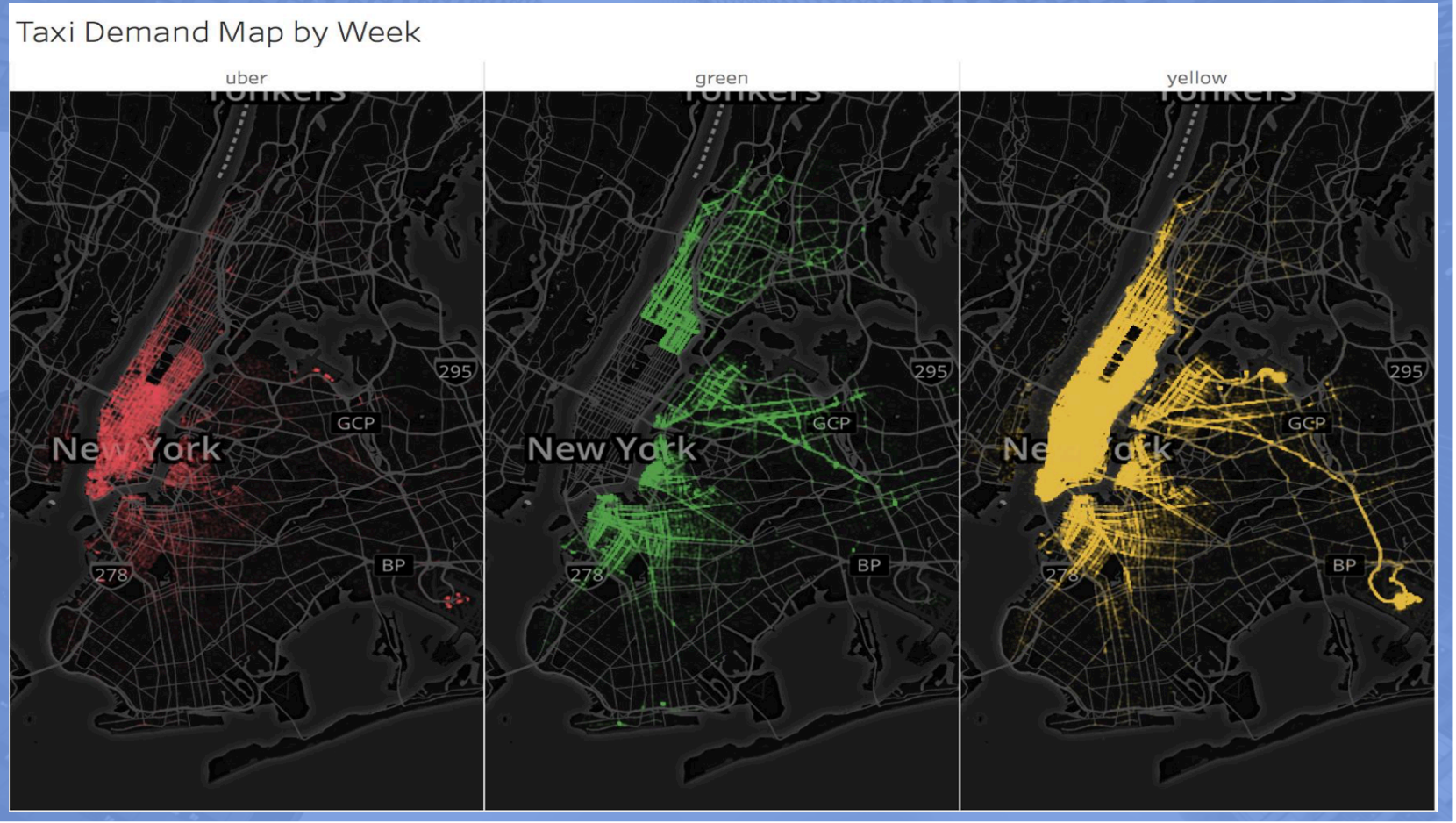
Data Sets ~ 71,000*5

Modeling Workflow

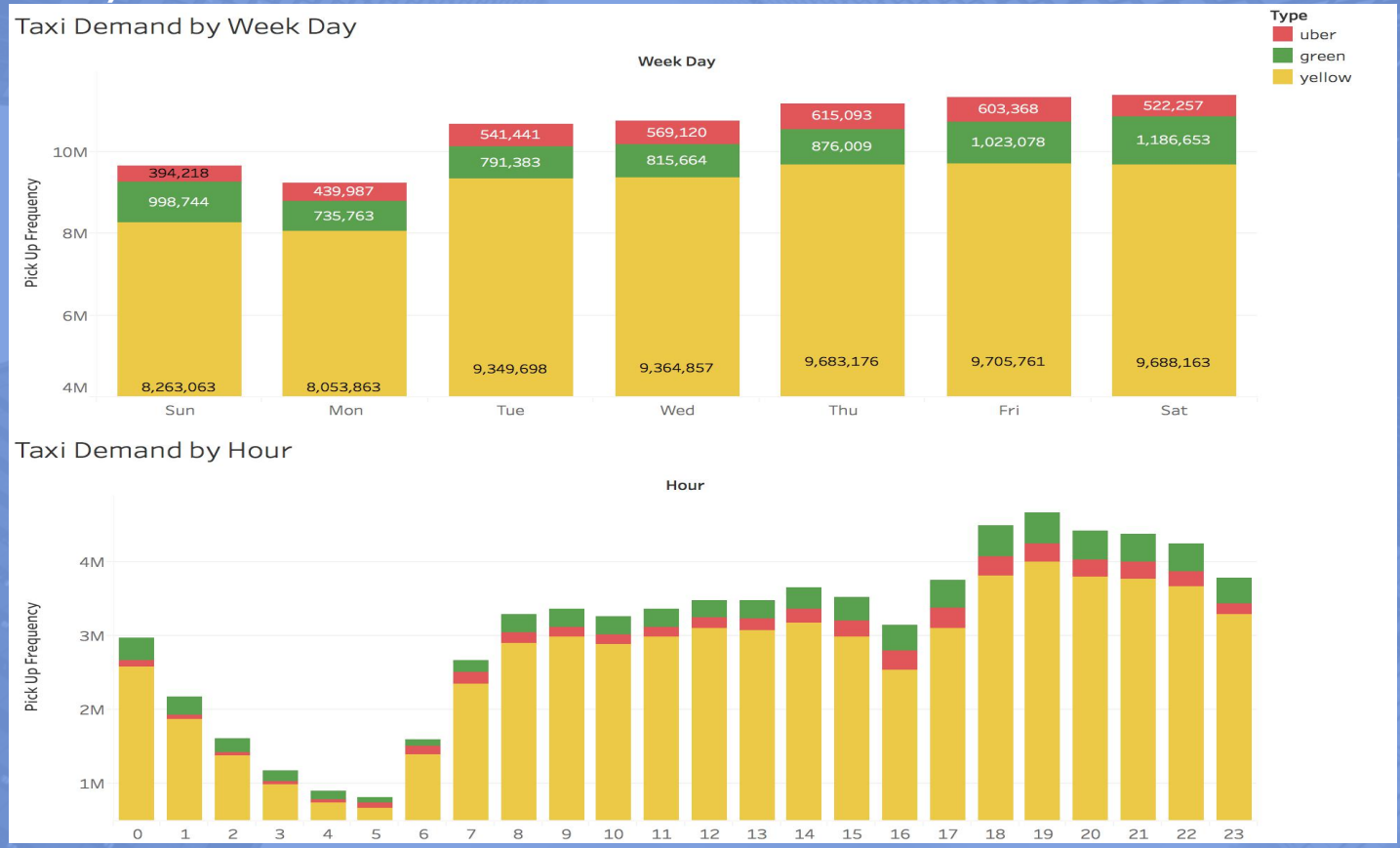


EDA (Tableau)

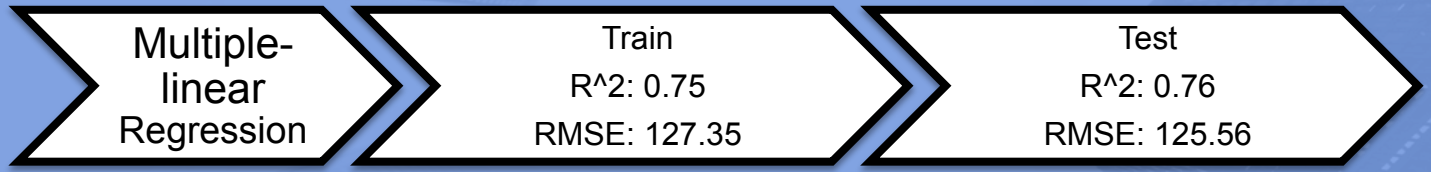

Weather
Hourly Taxi Demand by Temperature
Temp
Hourly Pick Up Frequency
25K
20K
15K
10K
33 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92
Hourly Taxi Demand by Rainfall
Rain
Hourly Pick Up Frequency
30K
25K
20K
15K
10K
5K
0K
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.11 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.27 0.28 0.32 0.42

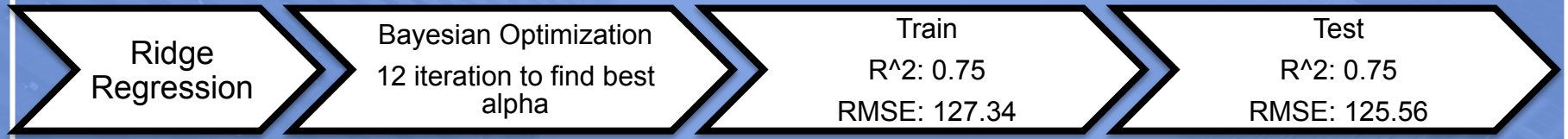

Taxi Demand Map by Week
uber
green
yellow
295
GCP
BP
278
New York

EDA (Tableau)

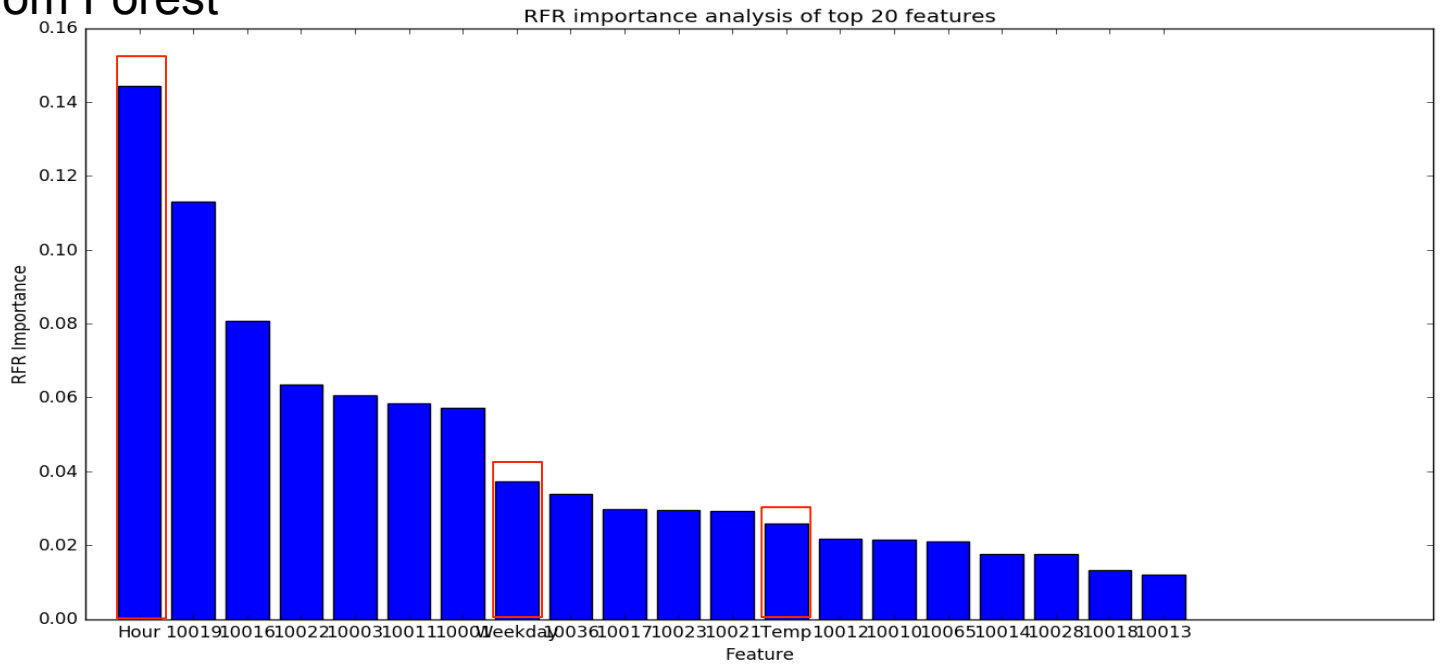

Taxi Demand by Week Day
Type
uber
green
yellow
Week Day
Pick Up Frequency
10M
8M
6M
4M
394,218
998,744
8,263,063
439,987
735,763
8,053,863
541,441
791,383
9,349,698
569,120
815,664
9,364,857
615,093
876,009
9,683,176
603,368
1,023,078
9,705,761
522,257
1,186,653
9,688,163
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Taxi Demand by Hour
Hour
Pick Up Frequency
4M
3M
2M
1M
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

MLR and Ridge Regression





Random Forest



1 Random Forests

Bayesian Optimization

2 iteration to find best max_features and n_estimators

Train

R^2: 0.99

RMSE: 16.05

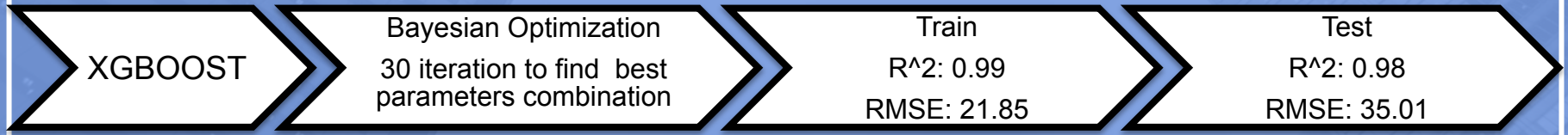
Test

R^2: 0.97

RMSE: 40.60

XGBOOST

Parameters	Range	Best (lowest MSE)
Max_depth	3 ~ 14	14
Learning_rate	0.01 ~ 0.2	0.1186
N_estimators	50, 1000	463
gamma	0.01 ~ 1.0	1.0
Min_child_weight	1 ~ 10	6.1929
Subsample	0.5 ~ 1	0.9675
Colsample_bytree	0.5 ~ 1	0.8544



Ensemble

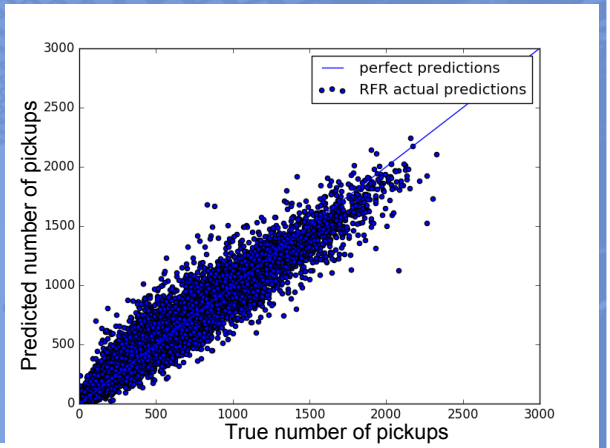
RandomForests Regression

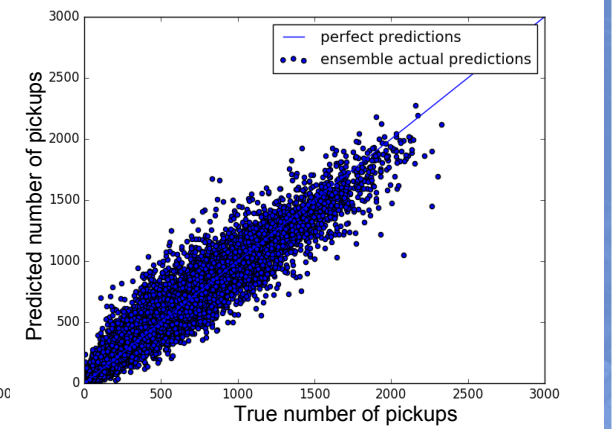
+

Xgboost Regression

→ Ensemble (Linear Regression)

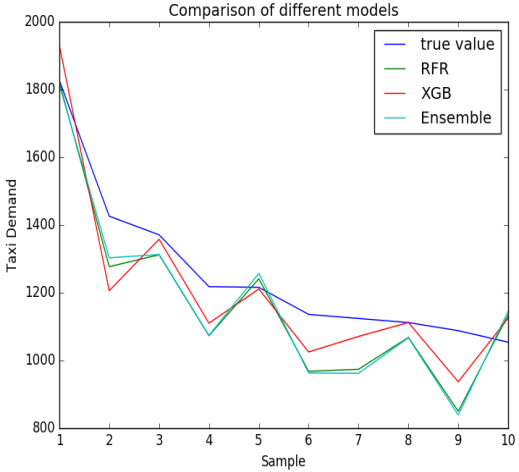
Model	R^2	RMSE
MLR	0.76	125.56
Ridge	0.75	125.56
RFR	0.97	40.06
XGBR	0.98	35.01
Ensemble	0.97	42.95

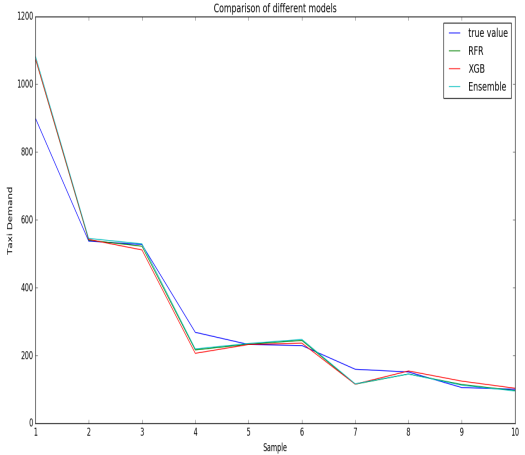


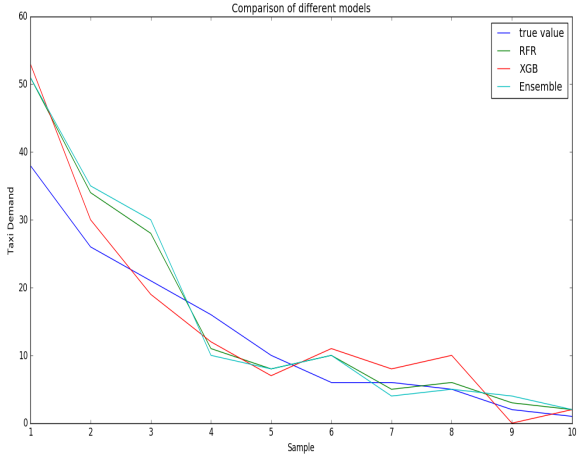


Further Comparison of Models

Subset	Demand	Size	RFR RMSE	XGBR RMSE	Ensemble RMSE
Subset 1	>=1000	2479	150	123	156
Subset 2	100 ~ 999	24759	75	66	80
Subset 3	<100	95472	8.7	7.2	9.5

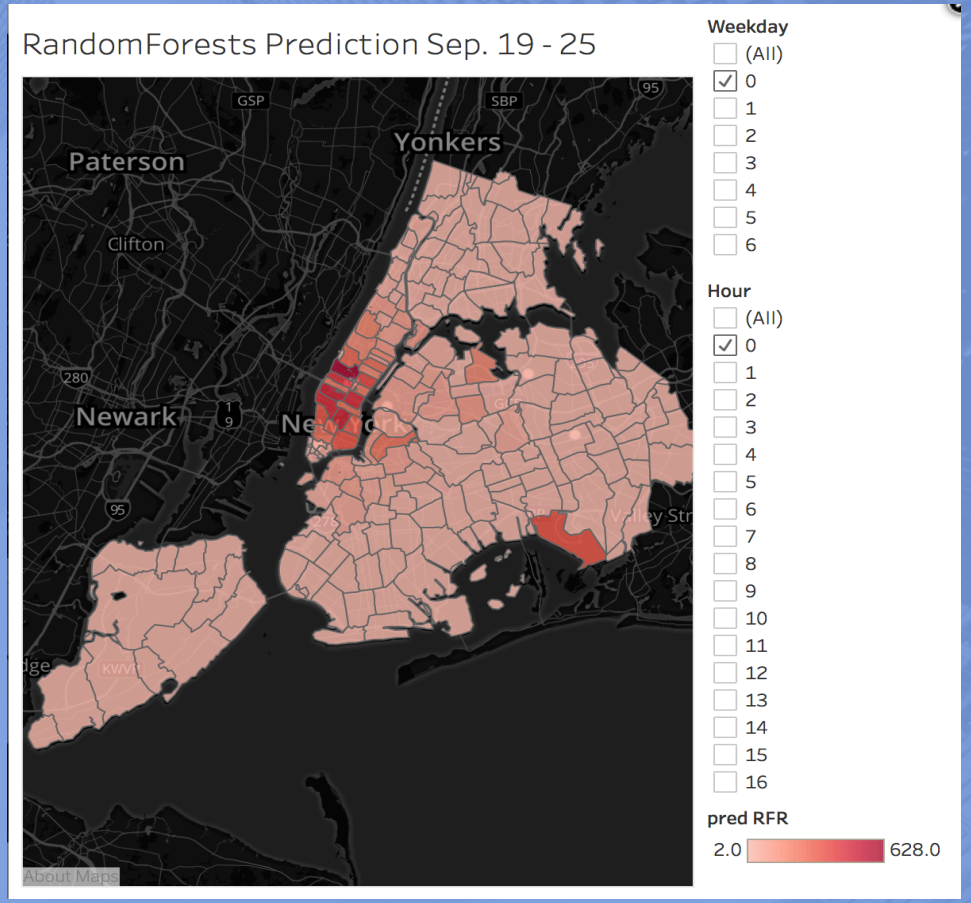






Prediction for the Coming Week

- RandomForest
- XGBOOST
- Ensemble

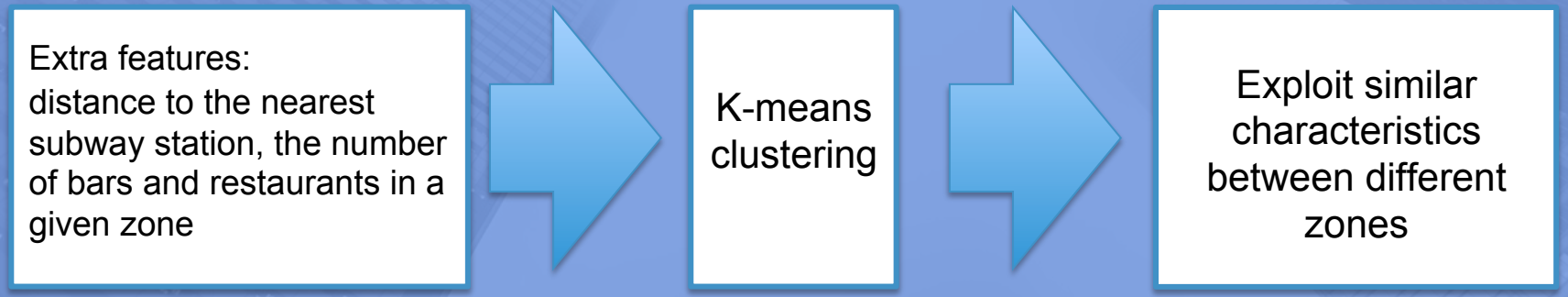


Conclusion & Future Work

- **Overall, our models for predicting taxi pickups in NYC performed well.**
- **XGBOOST performed best.**

Neutral network:

- **Automatically tune and model feature interactions**
- **Learn nonlinearities**




Thanks for your time
QUESTIONS?