



# 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 Preparation
- Modeling Workflow
- Exploratory Data Analysis
- Modelling
- Ensemble
- Prediction
- Conclusion and Future Work

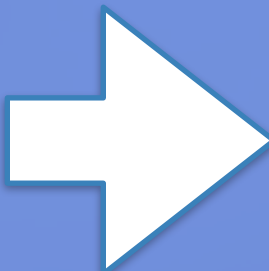
# Introduction

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



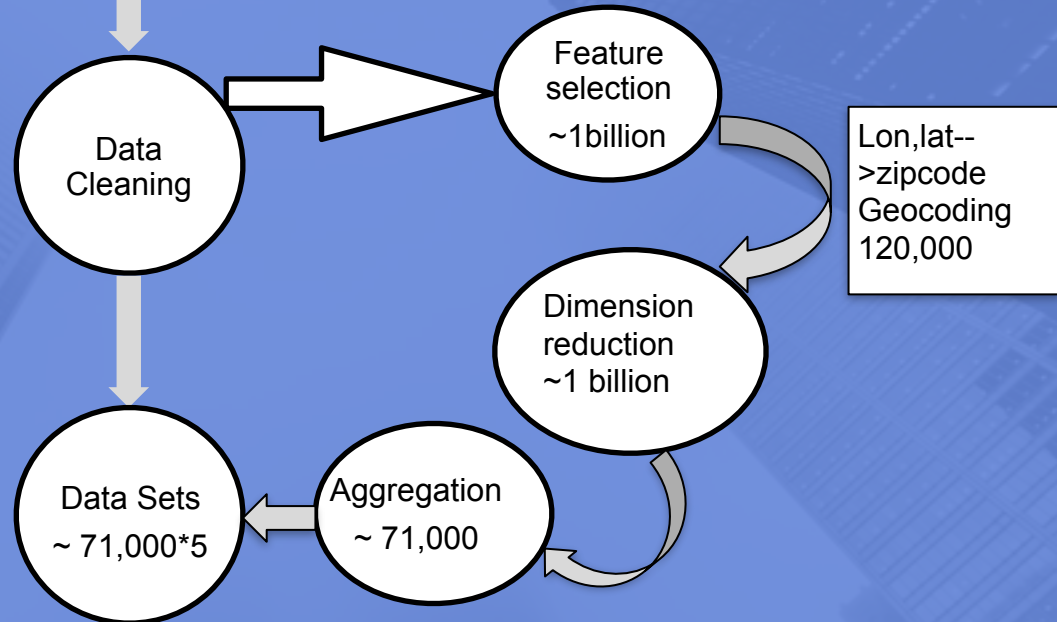
Output:

Predicted number of taxi pickups  
at the input time and location

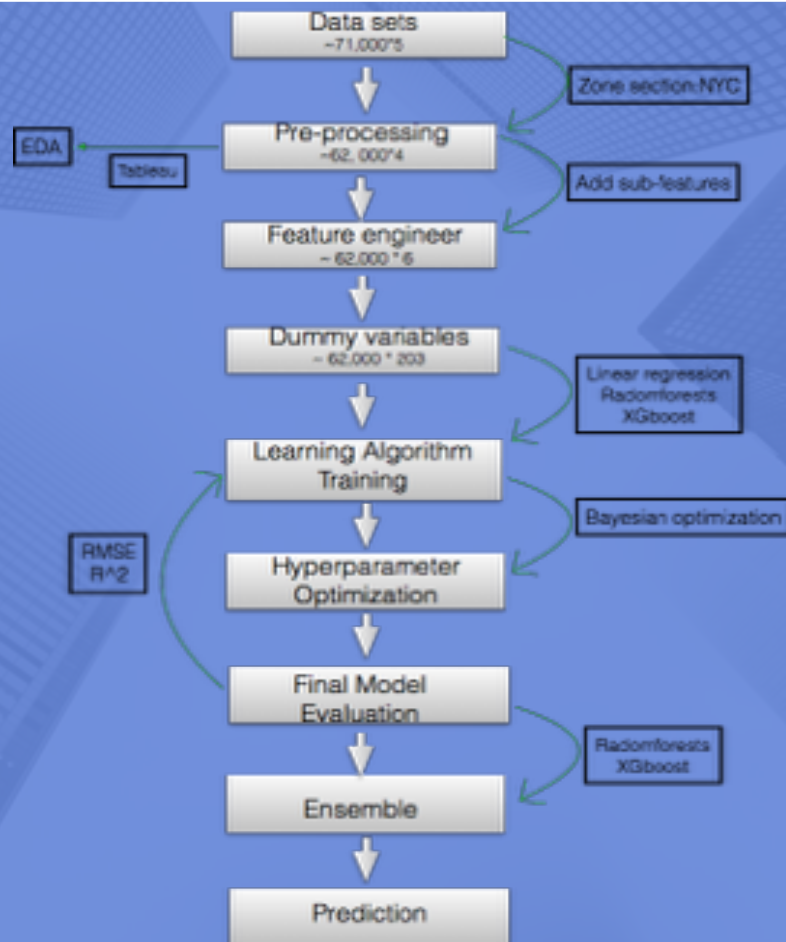


# Data Preparation

Raw Data			
Yellow 12.5G	Green 1.2G	Uber 0.22 G	Weather 0.001G



## Modeling Workflow

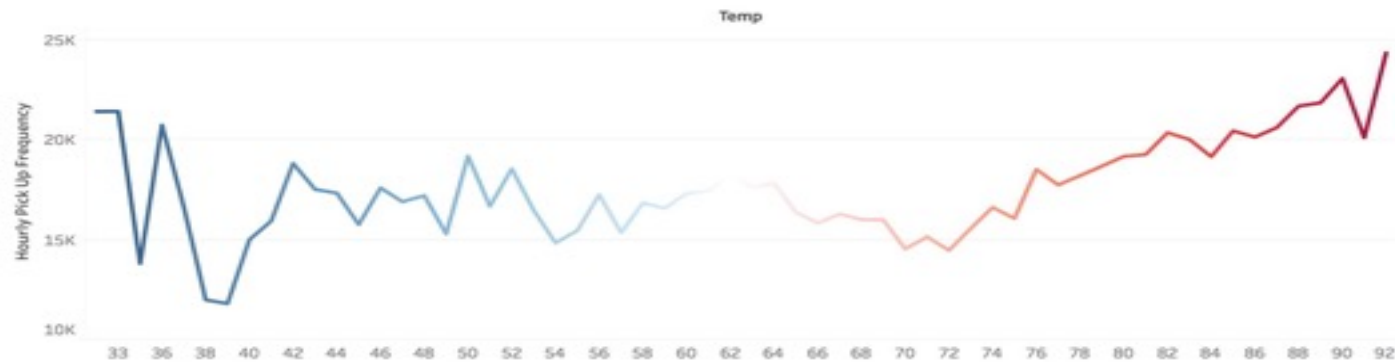


## EDA (Tableau)

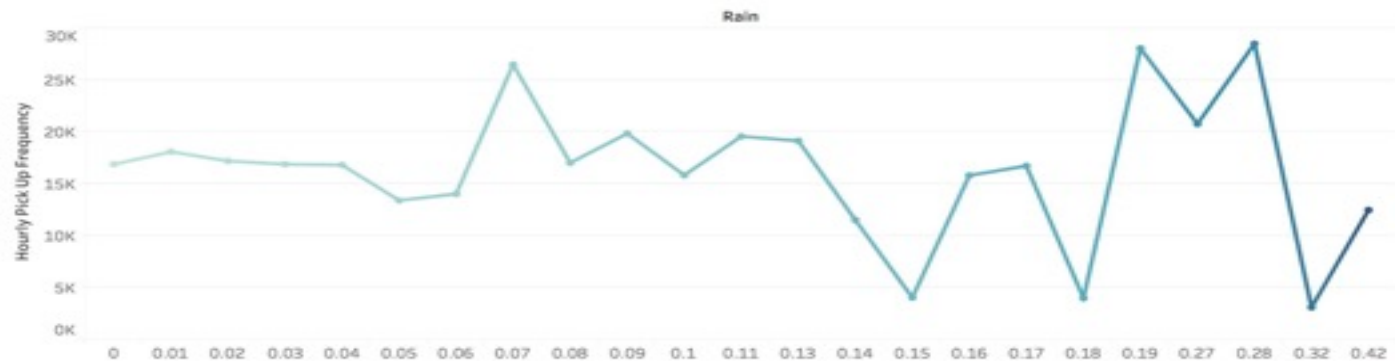


### Weather

#### Hourly Taxi Demand by Temperature

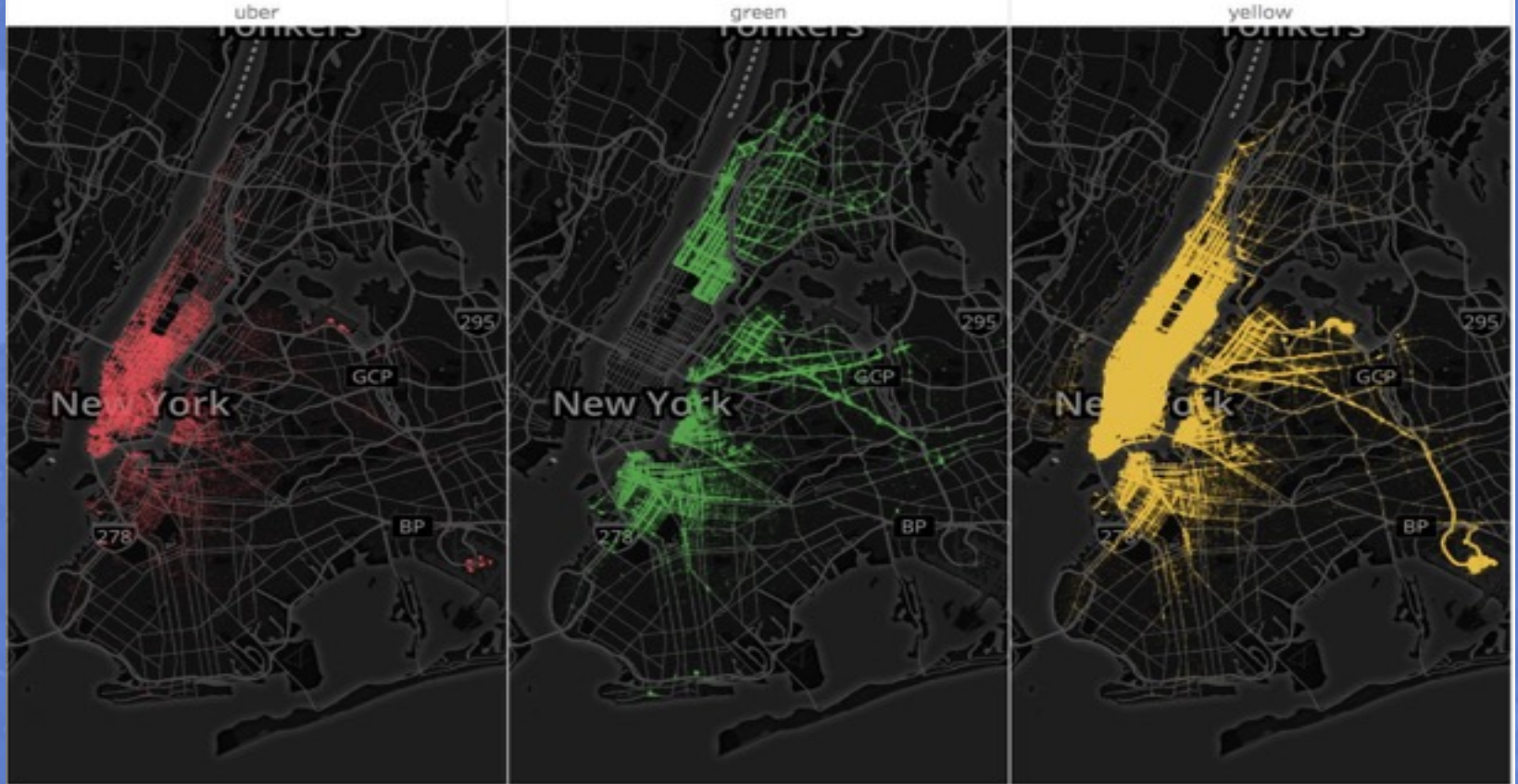


#### Hourly Taxi Demand by Rainfall





## Taxi Demand Map by Week





# EDA (Tableau)

Taxi Demand by Week Day



Taxi Demand by Hour



## MLR and Ridge

Multiple-linear  
Regression



Train  
 $R^2$ : 0.75  
RMSE: 127.35



Test  
 $R^2$ : 0.76  
RMSE: 125.56

Ridge Regression



Bayesian  
Optimization  
12 iteration to find  
best alpha

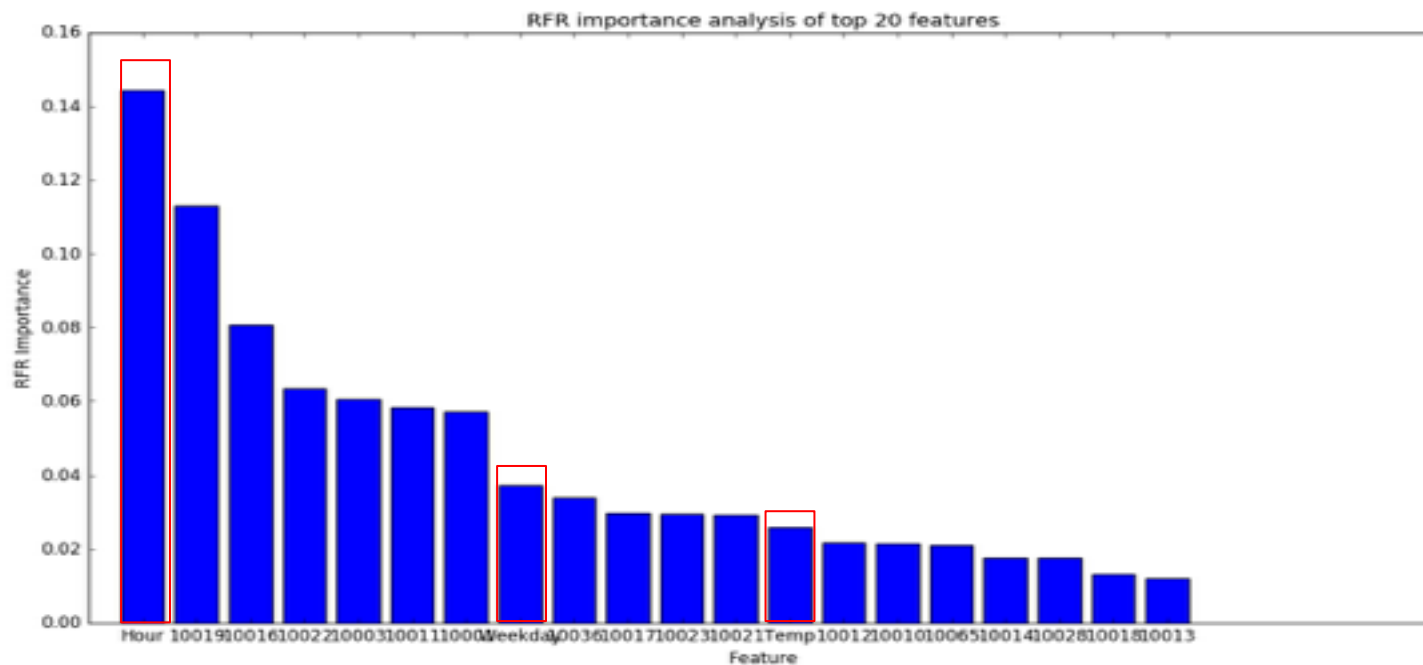


Train  
 $R^2$ : 0.75  
RMSE: 127.34



Test  
 $R^2$ : 0.75  
RMSE: 125.56

## Random Forests



RandomForest  
s



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

XGBOOST



Bayesian  
Optimization  
30 iteration to find  
best parameters  
combination



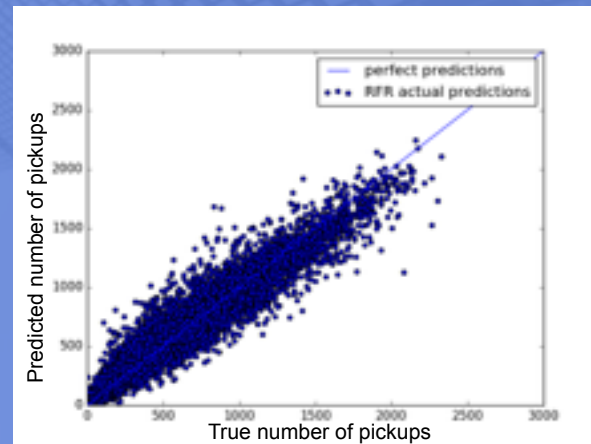
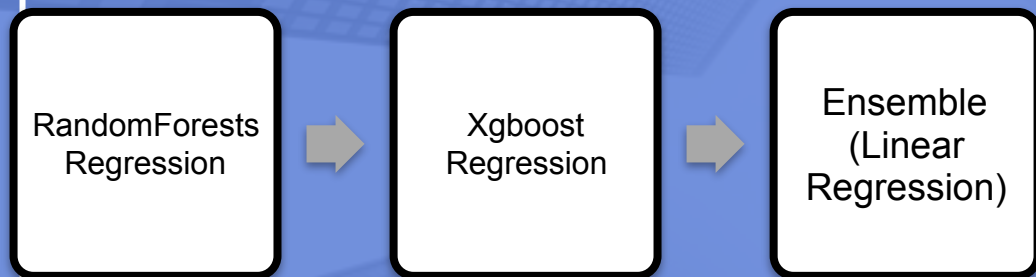
Train  
 $R^2$ : 0.99  
RMSE: 21.85



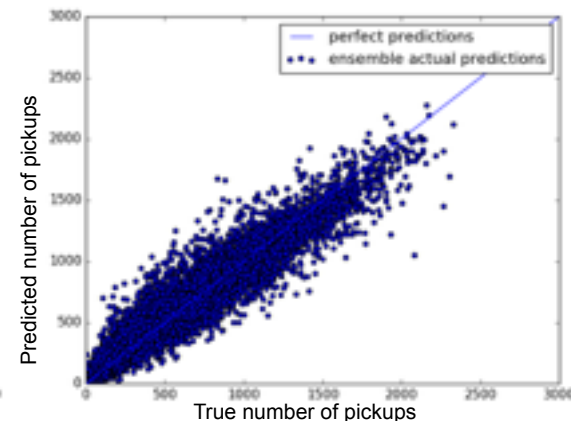
Test  
 $R^2$ : 0.98  
RMSE: 35.01



# Ensemble

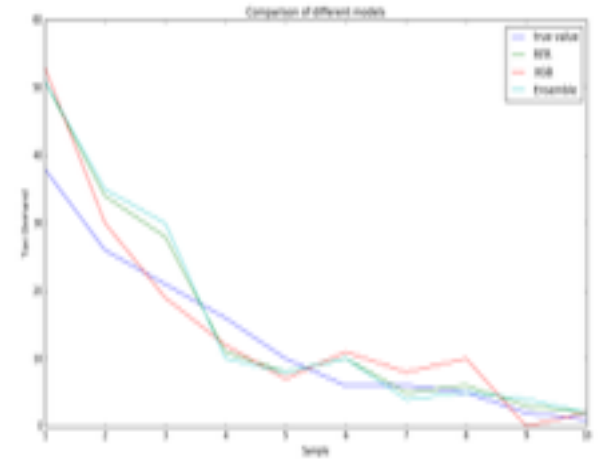
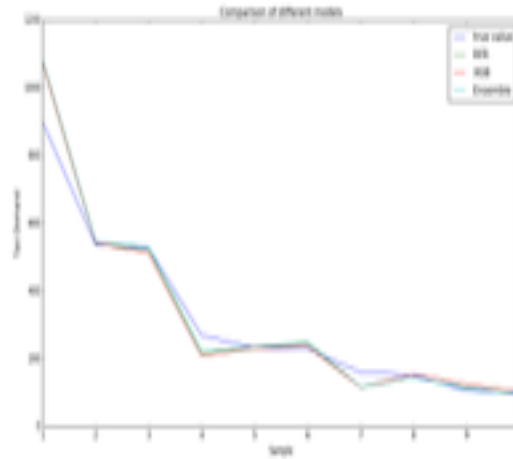
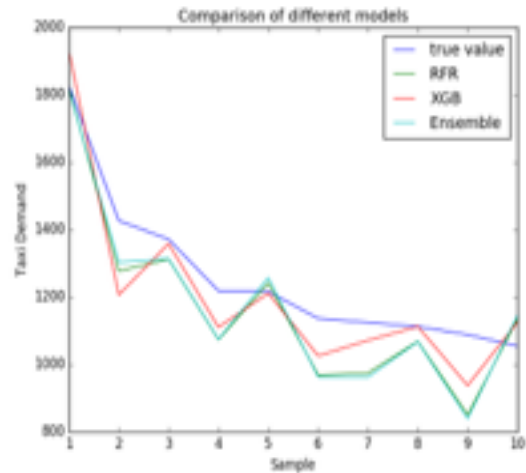


Model	R <sup>2</sup>	RMSE
MLR	0.76	125.56
Ridge	0.75	125.56
RFR	0.97	40.06
XGBR	0.98	35.01
Ensemble	0.97	42.95



## Further Comparison of Models

	Subset	Taxi Demand	Size	RFR RMSE	XGBR RMSE	Ensemble RMSE
	Subset 1	$\geq 1000$	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 and Future Work

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

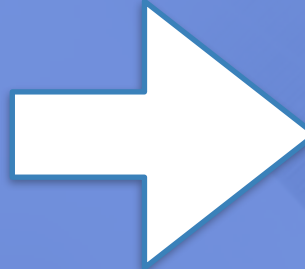
Neural network:

- automatically tune and model feature interactions
- learn nonlinearities

Extra features:  
distance to the nearest  
subway station, the number  
of bars and restaurants in a  
given zone



K-means  
clustering



Exploit similar  
characteristics  
between different  
zones





Thanks for your time  
**QUESTIONS?**

