San Francisco Bike Share Trip Duration Study

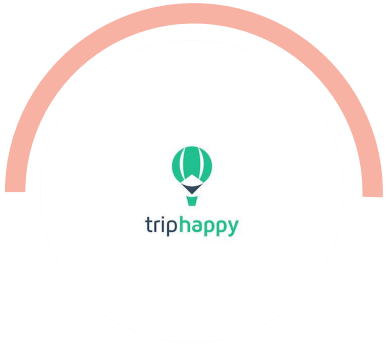
Group 20 **Small Face Manufacturer**

Yao Liu, Arthur Tian Qin, Vivian Chu Xiao, Todd Zhang, Reagan Yunzheng Zhao

Agenda

1. Dataset & Analytic Goals
2. Related Work
3. Preprocessing Algorithms & Efficiency
4. Machine Learning Outcome Comparison
5. Running Time Comparison
6. Conclusion & Lesson Learned

San Francisco Bay Area Bike Share Dataset (4GB)



Weather

3665 rows
24 columns
temperature, humidity, sea level pressure, visibility, wind speed

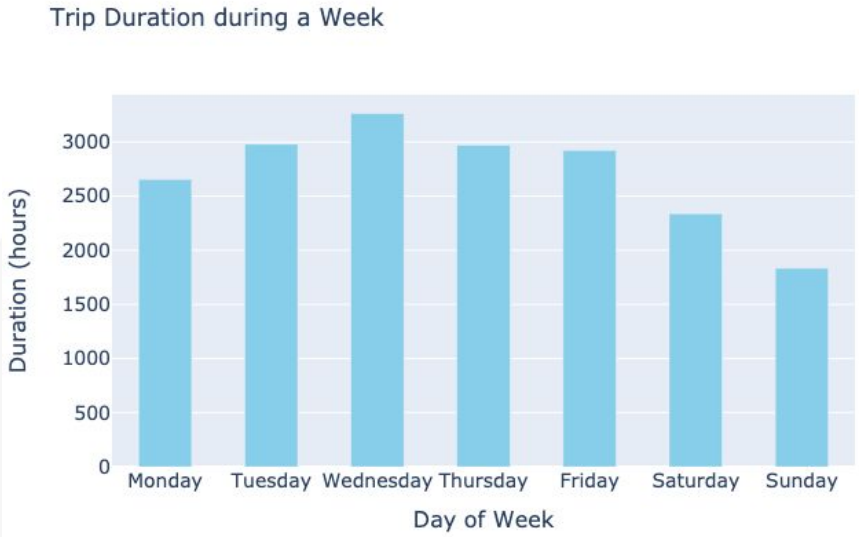
Trip

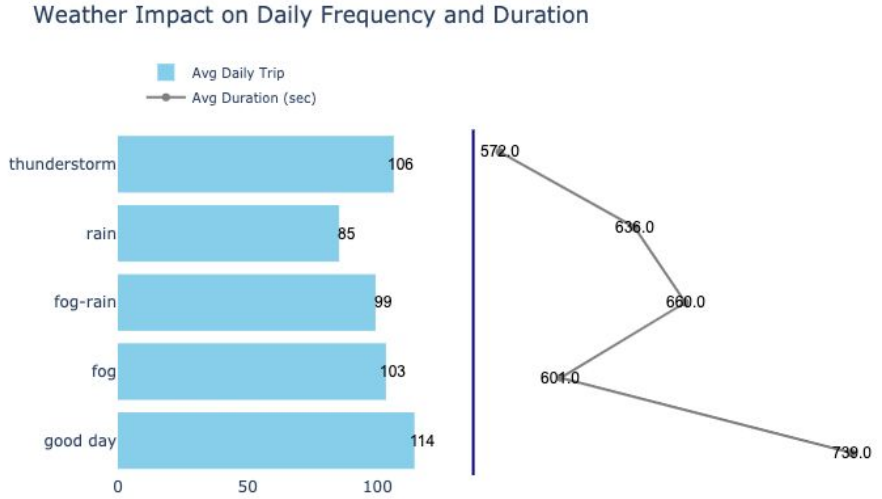
670000 rows
11 columns
Trip duration, subscription, start station, end station

Station

70 rows
7 columns
latitude and longitude

From our previous presentation





Analytic Goals

- Understand how features such as weather conditions affect trip duration
- Build models to predict trip durations based on bikeshare data
- Potentially help people optimize their transportation choice under different weather conditions

Related Works

1 Examine the effect of weather on bike share travel

2 Use feature importance to reduce features and choose features that can best predict duration

Modeling bike counts in a bike-sharing system considering the effect of weather conditions

Huthaifa I. Ashqar[a], Mohammed Elhenawy[b], Hesham A. Rakha[c,*]

[a] Booz Allen Hamilton, Washington, D.C., United States
[b] CARRS-Q, Queensland University of Technology, 130 Victoria Park Road, Kelvin Grove QLD 4059, Australia
[c] Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States

ARTICLE INFO

Keywords:
Bike counts prediction
Bike-sharing
Big data
Random Forest
Urban Computing

ABSTRACT

The paper develops a method that quantifies the effect of weather conditions on the prediction of bike station counts in the San Francisco Bay Area Bike Share System. The Random Forest technique was used to rank the predictors that were then used to develop a regression model using a guided forward step-wise regression approach. The Bayesian Information Criterion was used in the development and comparison of the various prediction models. We demonstrated that the proposed approach is promising to quantify the effect of various features on a large BSS and on each station in cases of large networks with big data. The results show that the time-of-the-day, temperature, and humidity level (which has not been studied before) are significant count predictors. It also shows that as weather variables are geographic location dependent and thus should be quantified before using them in modeling. Further, findings show that the number of available bikes at station i at time $t-1$ and time-of-the-day were the most significant variables in estimating the bike counts at station i.

1. Introduction

A growing population, with more people living in cities, has led to increased pollution, noise, congestion, and greenhouse gas emissions. One possible approach to mitigating these problems is encouraging the use of bikes through bike sharing systems (BSSs). BSSs are an important part of urban mobility in many cities and are sustainable and environmentally-friendly systems. As urban density and its related problems increase, it is likely that more BSSs will exist in the future. The relatively low capital and operational cost, ease of installation, use of advanced technologies for implementation and management, demonstrates a shift into the fourth generation of BSSs (Susan and Stacey, 2010).

In 2013, San Francisco launched the Bay Area BSS, a membership-based system providing 24 h a day, 7 days a week self-service access to short-term rental bicycles. Members can check out a bicycle from a network of automated stations, ride to the station nearest their destination, and leave the bicycle safely locked for someone else to use (Share, 2016). The Bay Area BSS is designed for short, quick trips, and as a result, additional fees apply to trips longer than 30 min. In this

Bike Share Travel Time Modeling: San Francisco Bay Area Case Study

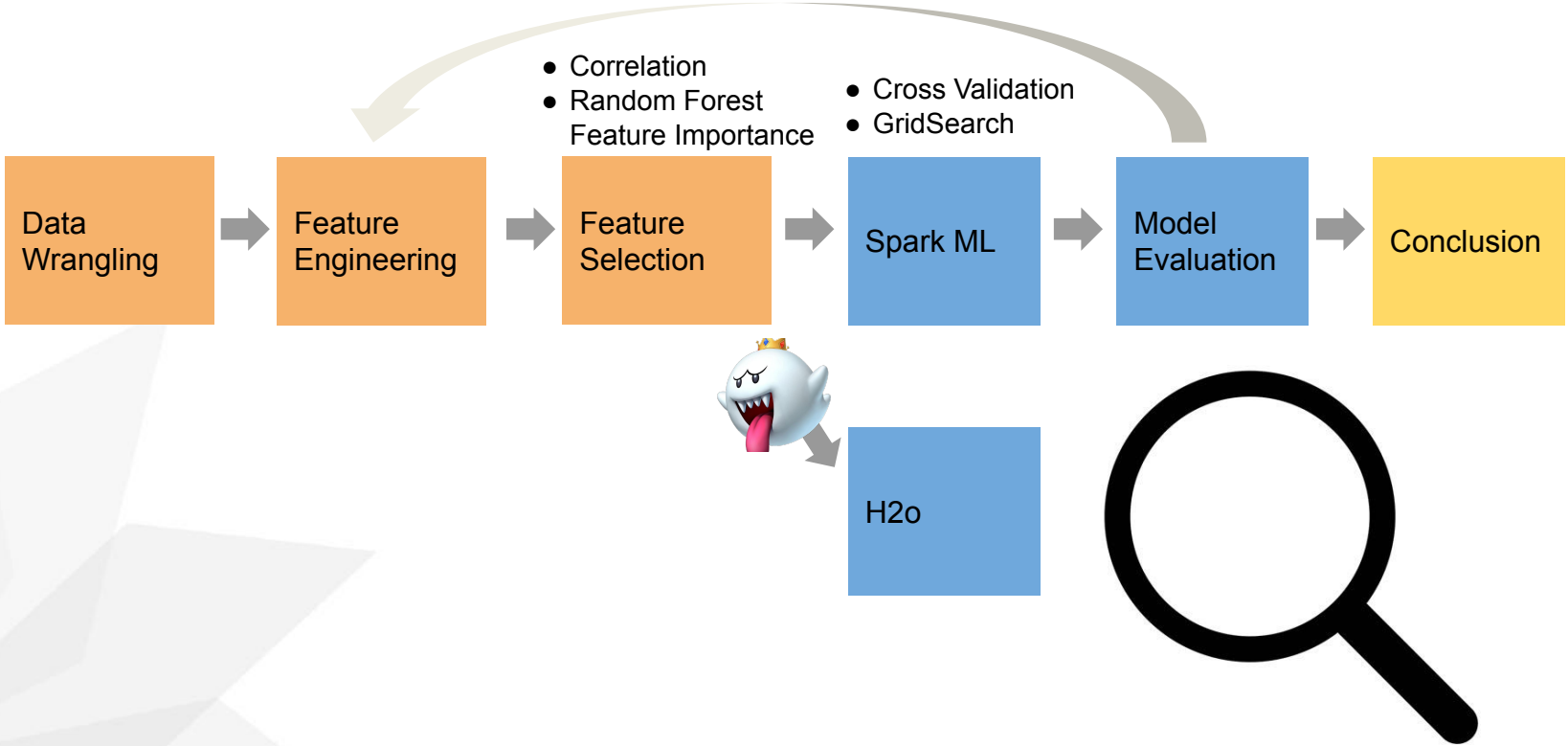
Ahmed Ghanem*, Mohammed Elhenawy*, Mohammed Almannaa§, Huthaifa I. Ashqar§ and Hesham A. Rakha*§
* Bradley Dept. of Electrical and Computer Engineering, Virginia Tech, Blacksburg, Virginia 24060
§ Charles E. Via, Jr. Dept. of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia 24060
aghanem, elhenawy, almannaa, hiashqar, hrakha@vt.edu

***Abstract*—Bike Share Systems (BSSs) are emerging in many US cities as a new sustainable transportation mode that provides a last-mile solution for short-distance transfers between different private and public transportation modes. In order to encourage the increased use of bikes as a mode of transportation, tools, measures, and planning techniques similar to those used for other transportation modes need to be developed. With precise information on the trip travel time, route planner systems can suggest optimal alternative routes, and manage and control traffic congestion. Although there is a growing body of literature dealing with BSSs, bike travel time has been studied sparingly up to this point. In this paper, we addressed this issue by developing different bike travel time models using random forest (RF), least square boosting (LSBoost) and artificial neural network (ANN) techniques. We studied 33 different predictors affecting bike travel time, including such predictors as travel distance, biker experience, time-of-day, and weather conditions. The RF model**

Roadway congestion levels began to rise again along with the US economy's recovery from the most recent recession. Congestion levels have not only returned to the pre-recession levels of 2000 and before, but they are now even greater, causing more congestion problems. By 2014, congestion had caused travel delay to increase to 6.9 billion hours per year, up from 5.2 billion hours per year in 2000. Additionally, congestion costs increased by nearly $46 billion between 2000 and 2014, reaching $160 billion in 2014 [4]. Ideally, the increased presence and use of BSSs will mean decreased congestion levels.

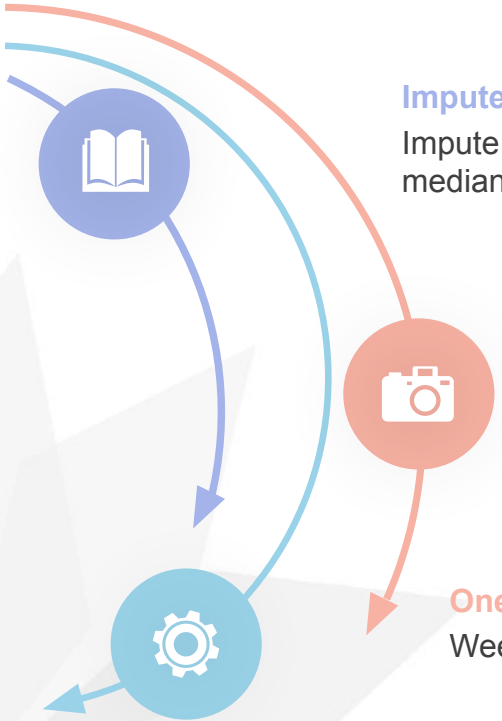
With growing warnings and worries about climate change and increased recommendations to reduce fossil fuel consumption, people are more open to using sustainable transportation

Our Approach



Preprocessing

Cluster specs *- m4.4xlarge/ 2 instances*
Runtime *- 96.25 seconds for Imputing, StringIndexing, and OneHotEncoding + 10.41 seconds for join*



Imputer

Impute missing values with the median/mode

```
from pyspark.ml.feature import Imputer
imputer = Imputer(inputCols=columns,
outputCols=imputed_columns).setStrategy("median")
mtrip_weather_imputed_part_one = imputer.fit(mtrip)
```

StringIndexer

Convert columns to numeric representations
Station name, subscription_type, precipitation_inches

```
from pyspark.ml.feature import StringIndexer
indexed_train_fit =
StringIndexer(inputCol=string_names_without_date[i],
outputCol=indexed_string_names_without_date[i]).set
HandleInvalid("keep").fit(mtrip_weather_imputed_fin)
```

OneHotEncoder

Weekday, month, hour

```
from pyspark.ml.feature import OneHotEncoder
onehotenc = OneHotEncoder(inputCol=c,
outputCol=c+"-onehot", dropLast=False)
newdf = onehotenc.transform(newdf).drop(c)
newdf = newdf.withColumnRenamed(c+"-onehot", c)
```

Feature Selection

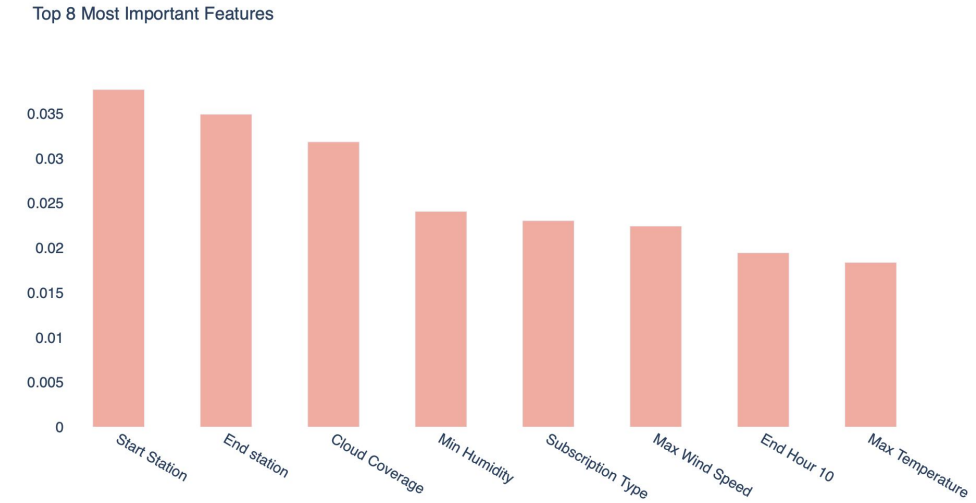
```
rf = RandomForestRegressor(maxDepth=10, maxBins= 100, minInstancesPerNode=1, minInfoGain = 0, seed=1)
rfmodel = rf.fit(f_train)
attrs = sorted((attr["idx"], attr["name"]) for attr in (chain(*lpoints_f_train
        .schema["features"]
        .metadata["ml_attr"]["attrs"].values())))

feature_scores = [(name, rfmodel.featureImportances[idx]) for idx, name in attrs if rfmodel.featureImportances[idx]
feature_scores = sorted(feature_scores, key=lambda tup: tup[1], reverse=True)
```

```
feature_scores
```

```
['zip_code', 'start_station_name', 'end_station_name', 'out_cloud_cover', 'out_min_humidity', 'subscription_type',
_wind_Speed_mph', 'start_name', 'out_max_temperature_f', 'end_name', 'out_mean_sea_level_pressure_inches', 'out_max
nt_f', 'out_wind_dir_degrees', 'out_mean_wind_speed_mph', 'out_mean_temperature_f', 'out_min_dew_point_f', 'out_max
eed_mph', 'out_max_sea_level_pressure_inches', 'start_long', 'start_id', 'start_weekday', 'start_hour', 'start_mont
 weekday', 'end_hour', 'end_month', 'label']
```

Feature Importance



Model Performance

01 **Linear Regression**

RMSE = 1161

cross validated

02 **Regularized Linear Regression**

RMSE = 1161

cross validated

03 **Random Forest**

RMSE = 1154.6

randomly and manually validated

04 **Decision Tree**

RMSE = 1346.7

randomly and manually validated

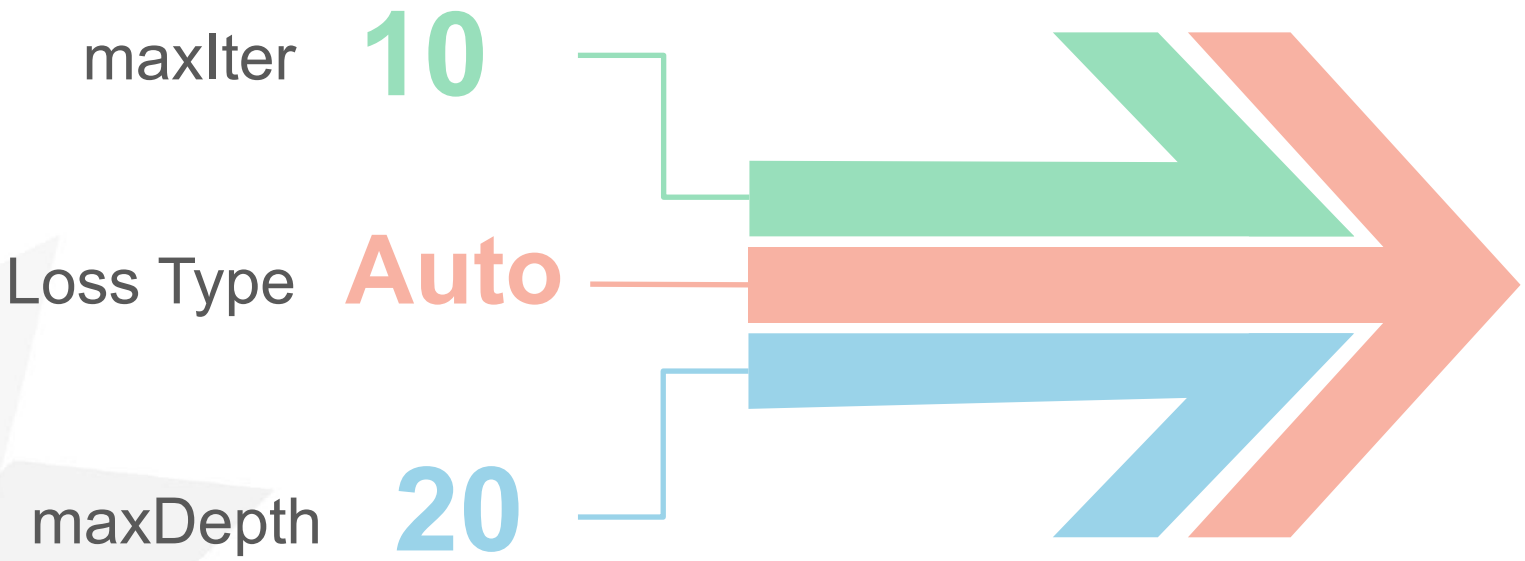
05 **Gradient Boosting**

RMSE = 1143.8

randomly and manually validated

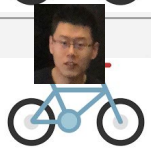
Cluster specs *- m4.4xlarge/ 2 instances*

Gradient Boosting Machine is our best model



Run Time of Best Model

m - general purpose, r - memory optimized. i - storage optimized

	Instance Type	Instance Size	# of Instances	Run Time
	m5	8xlarge	4	44.80s
	r4	8xlarge	4	39.55s
	m5a	8xlarge	4	50.90s
	r5a	8xlarge	4	43.34s
	i2	8xlarge	4	50.51s

Lessons Learned

01 Gradient Boost outperformed the other models (Linear Regressions, Random Forest, XGBoost and Gradient Boost) for travel time prediction

02 Noises significantly impact results. By dropping misleading outliers, RMSE dropped over 75%. Consider changing evaluation metrics.

03 Cross Validation on tree-based Spark ML takes long time, potentially due to how Spark ML optimizes splits of continuous variables (approximating quantile instead of using exact splits)

04 For Bike Riders : Optimize path according to duration
For Bike Sharing Service Company: maintain the system in a balanced state (optimize docks for different stations)

Thank you