SCALABILITY

Sankarshan and Paul
Airbnb Prediction Project

MEANING OF SCALABILITY FOR OUR PROJECT

- Built a model pipeline in Spark that instantiates every possible regression model based on parameter inputs and combinations of input features

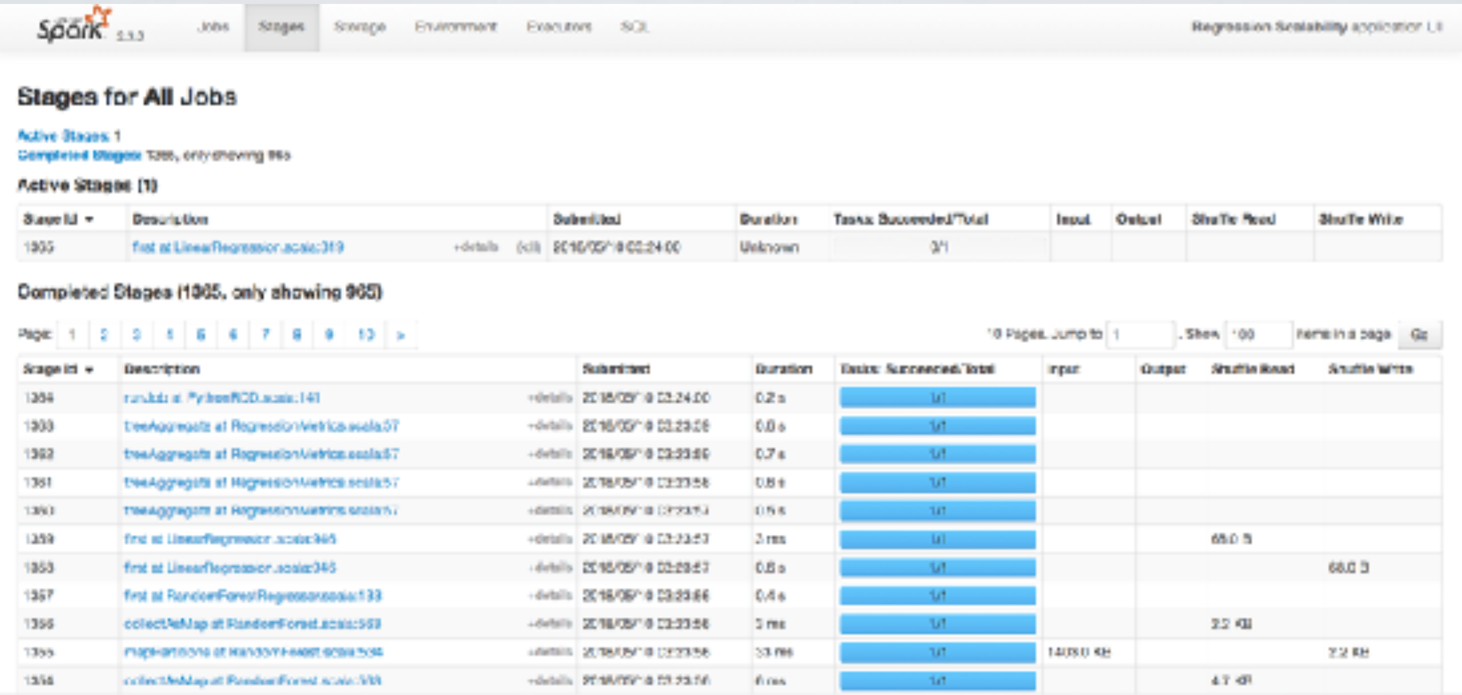
- Pipeline: Model Instantiation, Feature Importance, Evaluation Metrics

- Goal: find the best model features with brute force - or at least get an idea of where to look!

PERFORMANCE

- We have only run our code in Spark on a personal computer - AWS coming soon!

- Works for as many as 9 features (1.5 hours), breaks with 10 or more.



Stages for All Jobs

Active Stages: 1
Completed Stages: 1365, only showing 965

Active Stages (1)

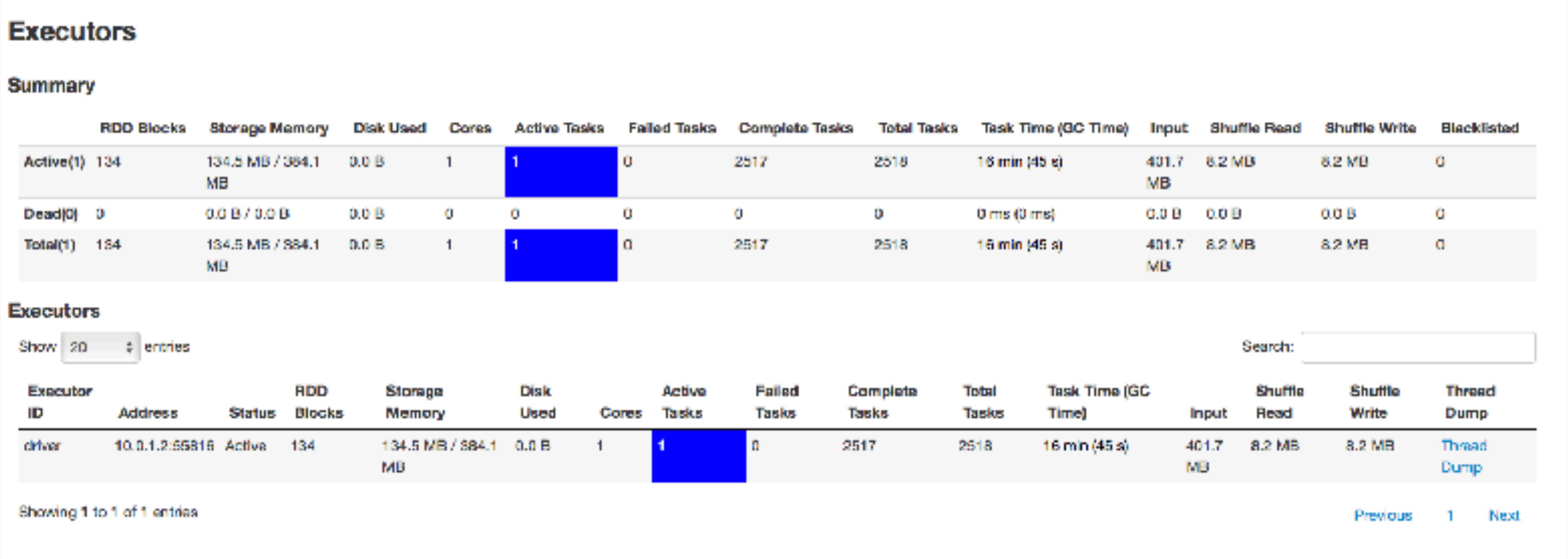
Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1365	first at LinearRegression.scala:319 +details (kill)	2018/05/18 03:24:00	Unknown	0/1				

Completed Stages (1365, only showing 965)

Page: 1 2 3 4 5 6 7 8 9 10 >

10 Pages. Jump to 1 . Show 100 items in a page Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1364	runJob at PythonRDD.scala:141 +details	2018/05/18 03:24:00	0.2 s	1/1				
1363	treeAggregate at RegressionMetrics.scala:57 +details	2018/05/18 03:23:59	0.0 s	1/1				
1362	treeAggregate at RegressionMetrics.scala:57 +details	2018/05/18 03:23:59	0.7 s	1/1				
1361	treeAggregate at RegressionMetrics.scala:57 +details	2018/05/18 03:23:58	0.8 s	1/1				
1360	treeAggregate at RegressionMetrics.scala:57 +details	2018/05/18 03:23:58	0.5 s	1/1				
1359	first at LinearRegression.scala:345 +details	2018/05/18 03:23:57	3 ms	1/1			68.0 B	
1358	first at LinearRegression.scala:345 +details	2018/05/18 03:23:57	0.8 s	1/1				68.0 B
1357	first at RandomForestRegressor.scala:133 +details	2018/05/18 03:23:56	0.4 s	1/1				
1356	collectAsMap at RandomForest.scala:563 +details	2018/05/18 03:23:56	3 ms	1/1			2.2 KB	
1355	mapPartitions at RandomForest.scala:534 +details	2018/05/18 03:23:56	33 ms	1/1	14080 KB			2.2 KB
1354	collectAsMap at RandomForest.scala:553 +details	2018/05/18 03:23:56	6 ms	1/1			4.7 KB	



Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Blacklisted
Active(1)	134	134.5 MB / 384.1 MB	0.0 B	1	1	0	2517	2518	16 min (45 s)	401.7 MB	8.2 MB	8.2 MB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	134	134.5 MB / 384.1 MB	0.0 B	1	1	0	2517	2518	16 min (45 s)	401.7 MB	8.2 MB	8.2 MB	0

Executors

Show 20 entries

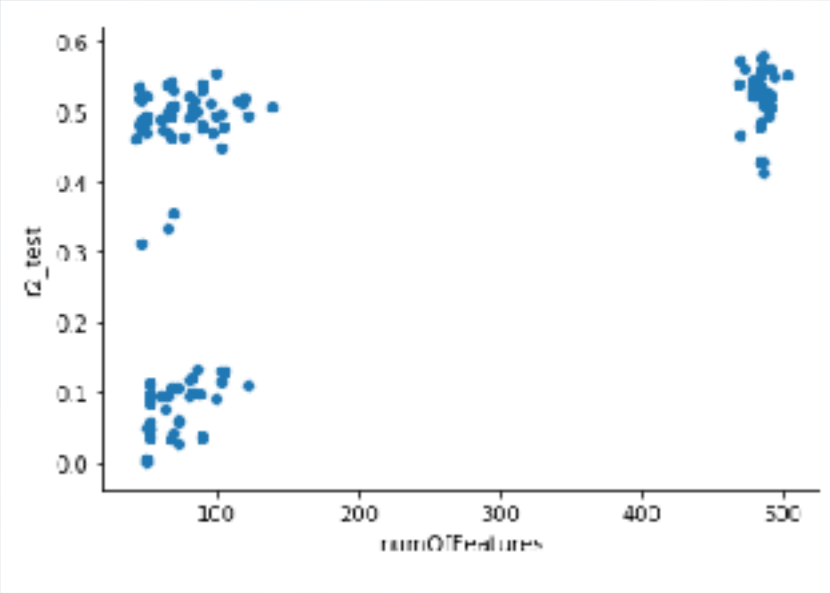
Search:

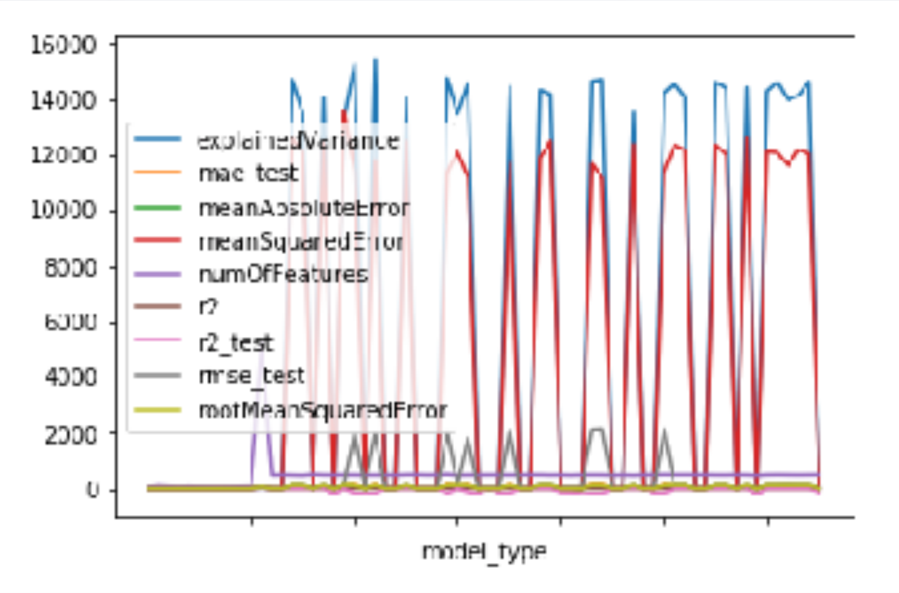
Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	10.0.1.2:55815	Active	134	134.5 MB / 384.1 MB	0.0 B	1	1	0	2517	2518	16 min (45 s)	401.7 MB	8.2 MB	8.2 MB	Thread Dump

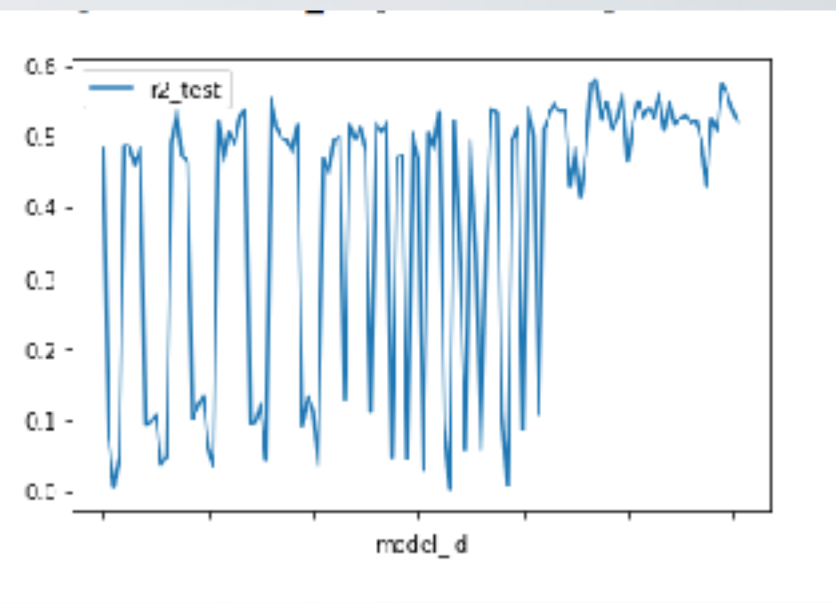
Showing 1 to 1 of 1 entries

Previous 1 Next

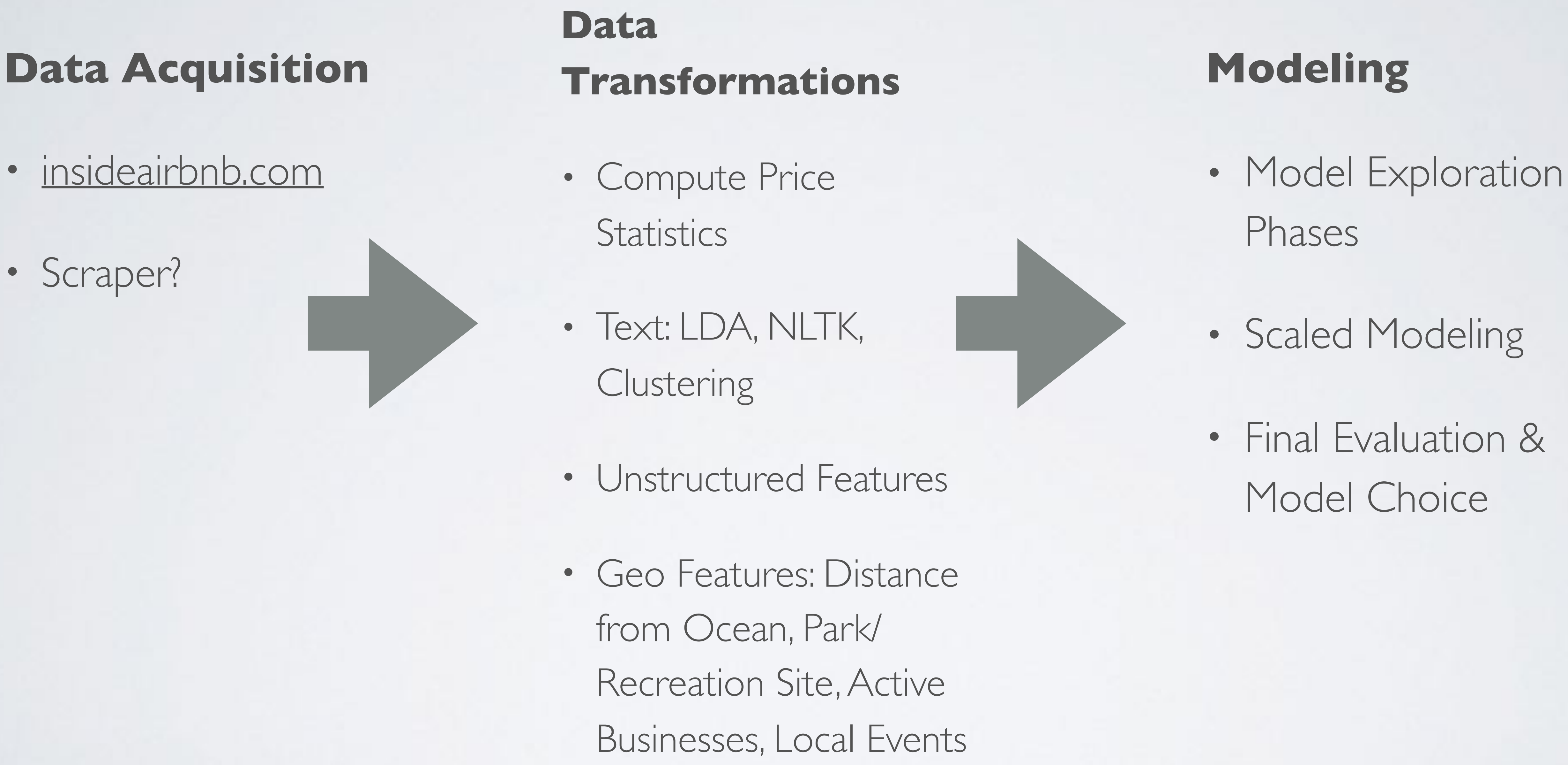
RESULTS


0.6
0.5
0.4
0.3
0.2
0.1
0.0
r2_test
100
200
300
400
500
numOfFeatures


16000
14000
12000
10000
8000
6000
4000
2000
0
explainedVariance
mae_test
meanAbsoluteError
meanSquaredError
numOfFeatures
r2
r2_test
mse_test
rootMeanSquaredError
model_type


0.6
0.5
0.4
0.3
0.2
0.1
0.0
r2_test
model_id

DATA PIPELINE



Data Acquisition

- insideairbnb.com
- Scraper?

Data Transformations

- Compute Price Statistics
- Text: LDA, NLTK, Clustering
- Unstructured Features
- Geo Features: Distance from Ocean, Park/ Recreation Site, Active Businesses, Local Events

Modeling

- Model Exploration Phases
- Scaled Modeling
- Final Evaluation & Model Choice

NEXT STEPS

- Run in AWS
- Expand capabilities of scalable pipeline
- Conclude work on scraper