CS 66 Final Project:
Dataset Size vs Testing Accuracy

Luke Pietrantonio

Question

1. How does dataset size affect testing accuracies?
    a. How does this change from model to model?
        i. Standard Linear Regression
        ii. Random Forest Linear Regression
        iii. Logistic Regression

Question cont.

Why is this important?

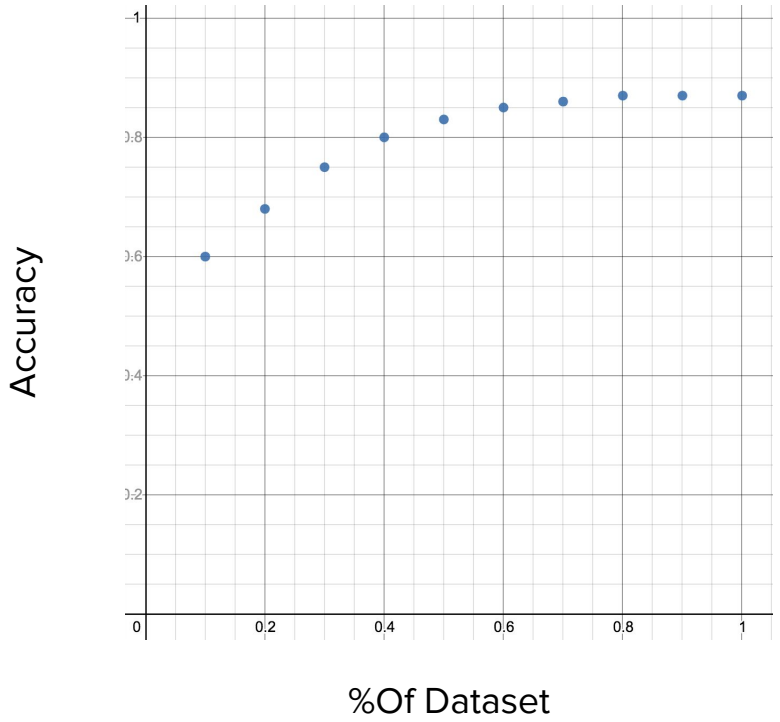
Can you achieve near perfect accuracies by increasing dataset size?

?????

Hypothesis

- As dataset size increases, so will model accuracy/score.
- But eventually plateau

Hypothesised Behavior:
%Of Dataset vs Accuracy



Dataset

- Bike Sharing Dataset
  - Hadi Fanaee T of The Laboratory of Artificial Intelligence and Decision Support
  - Collected From Capital Bikeshare in Washington D.C.
    - January 1st, 2011 to December 31st 2012
  - 17,379 unique entries
  - 14 Features

(Capital Bikeshare)

Dataset cont.

capital bikeshare

(Capital Bikeshare)

+

(Dreamstime)

Methods

SkLearn

- Linear Regression
- Random Forest Regression
- Logistic Regression

Methods cont.

Data Prep for Both

- Choose best features/remove repetitive features
- Split up data into Divisions

Data Prep for Linear Regression

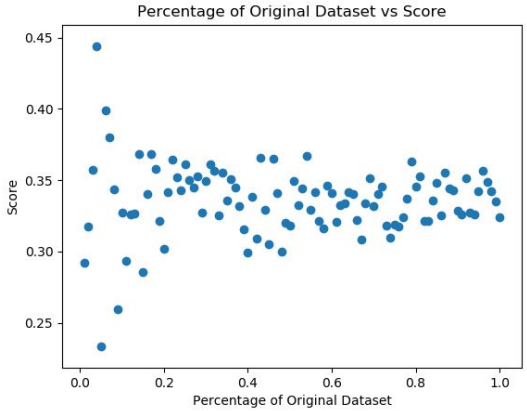
- No Additional Prep

Data Prep for Logistic Regression

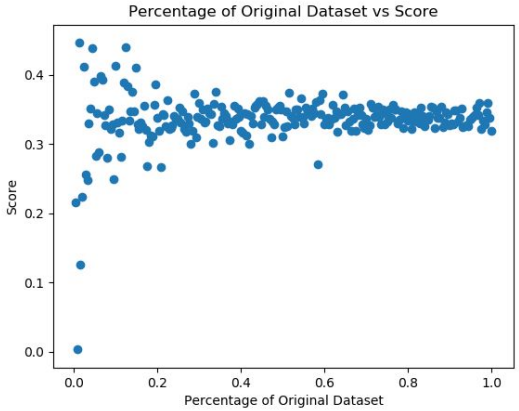
- Outputs: Continuous ➜ Categorical

(Shutterstock)

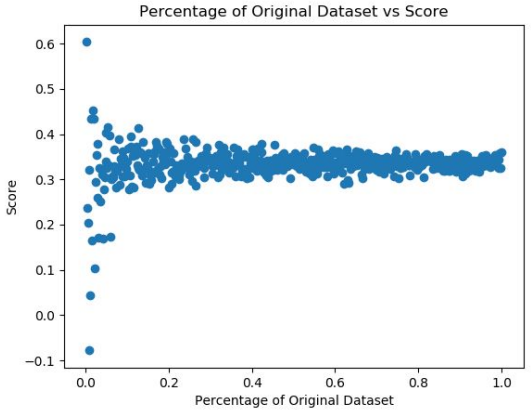
Results: Linear Regression



100 Divisions

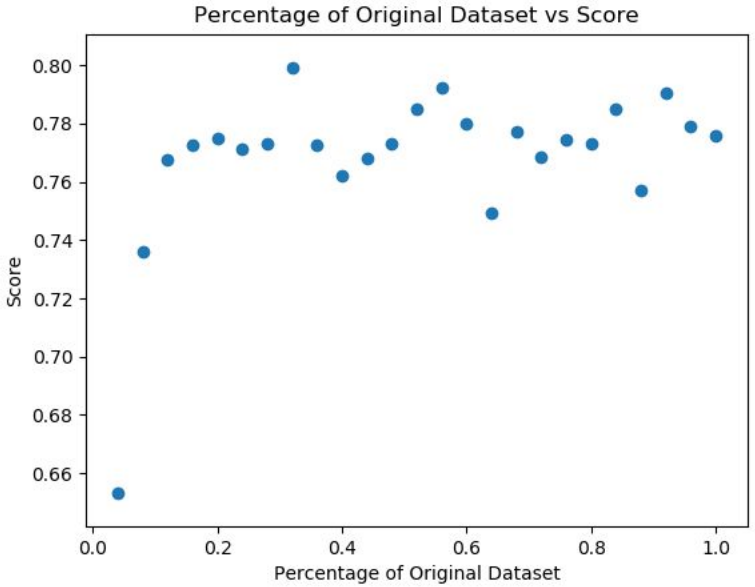


250 Divisions

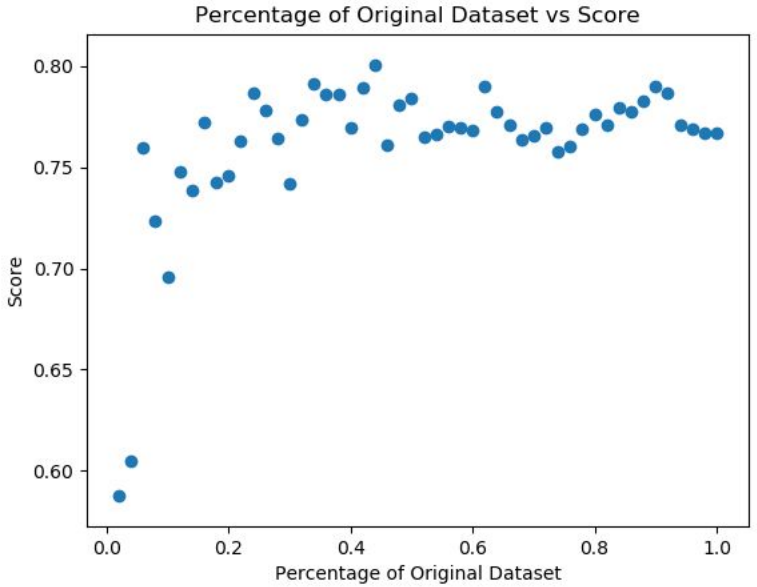


500 Divisions

Results: Random Forests

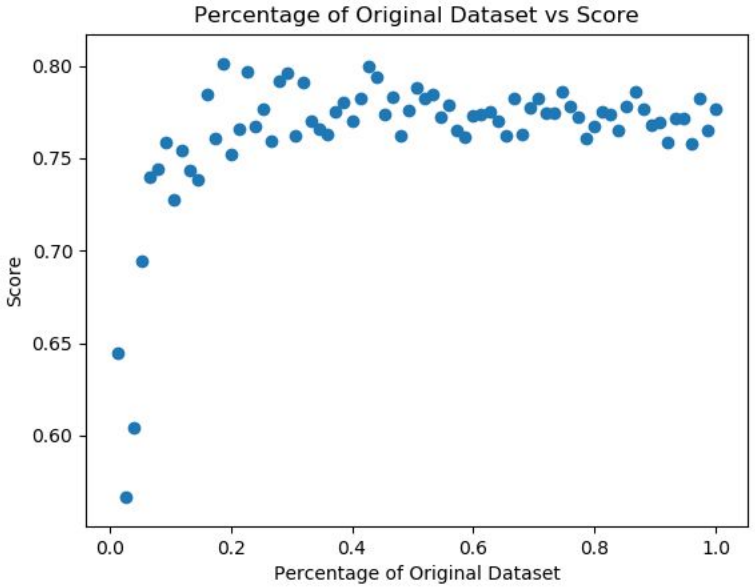


25 Divisions

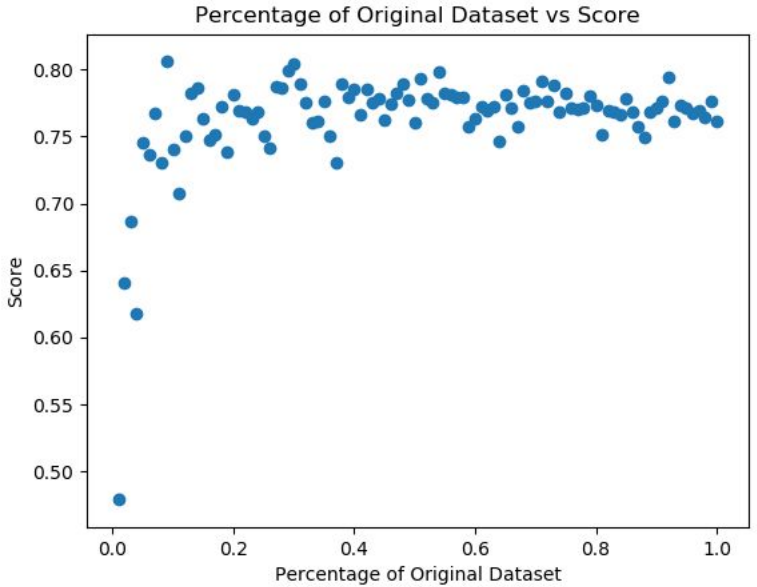


50 Divisions

Results: Random Forests cont.

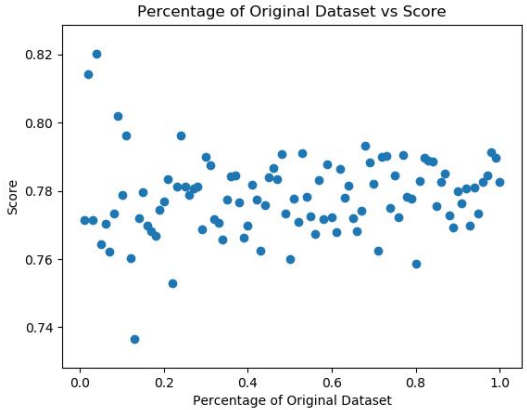


75 Divisions

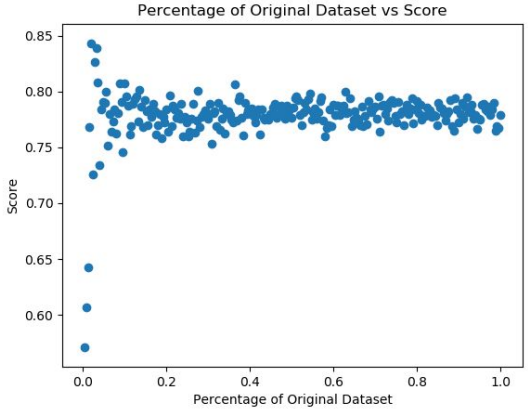


100 Divisions

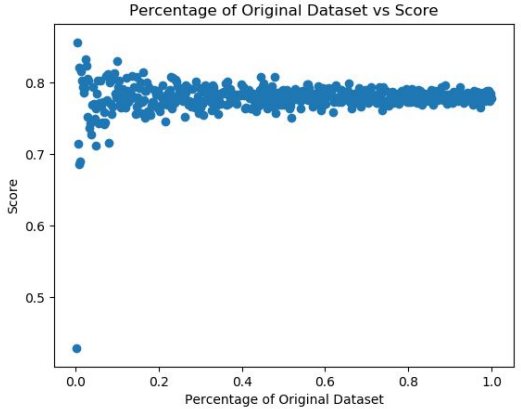
Results: Logistic Regression



100 Divisions

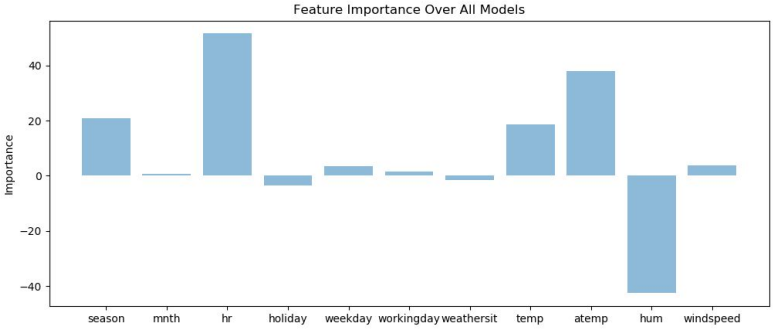


250 Divisions

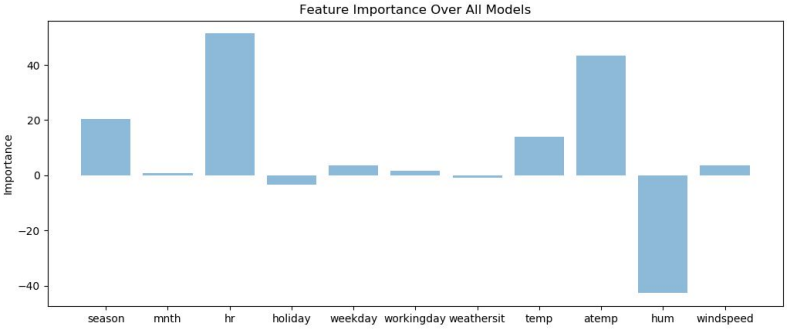


500 Divisions

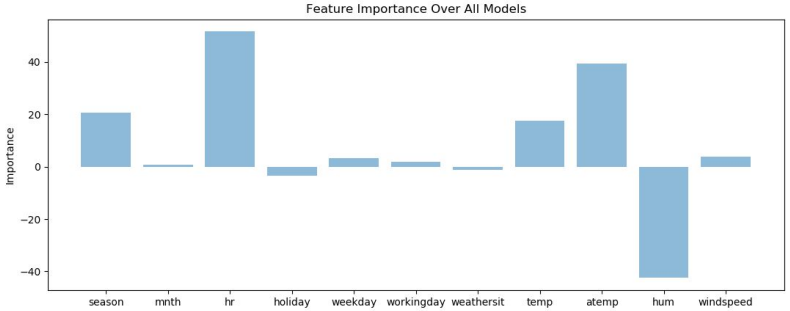
Feature Analysis: Linear Regression



100 Divisions



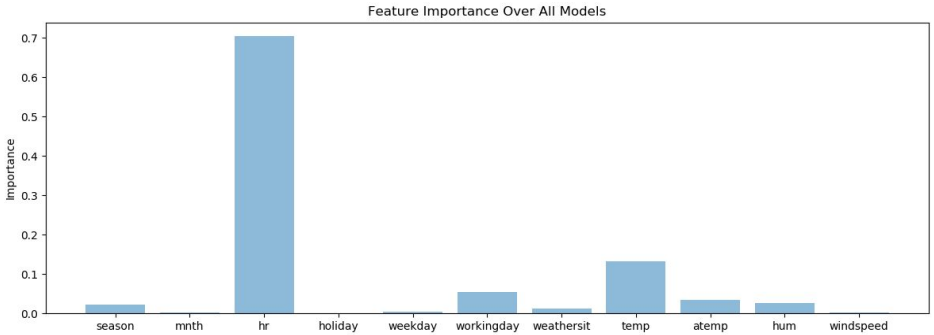
250 Divisions



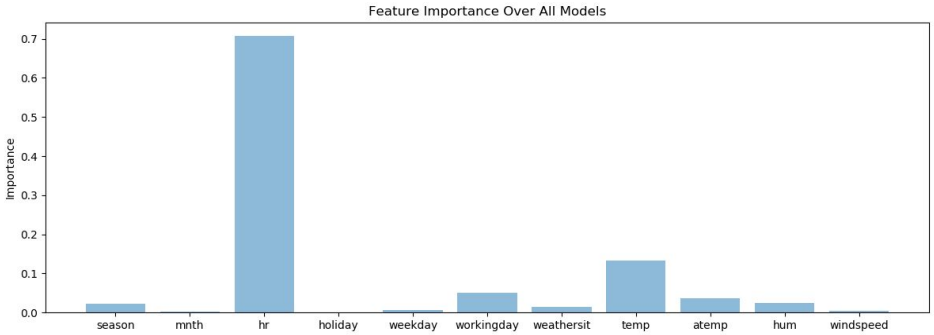
500 Divisions

Feature Analysis: Random Forests

25 Divisions

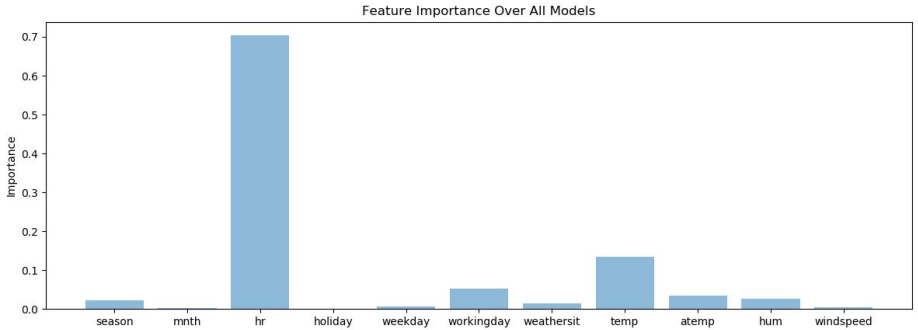


50 Divisions

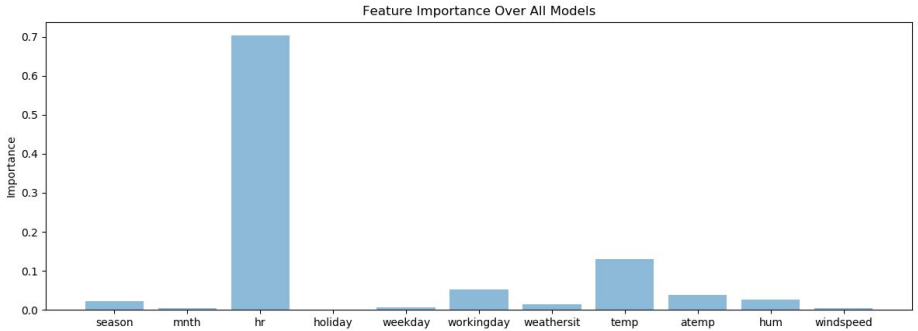


Feature Analysis: Random Forests

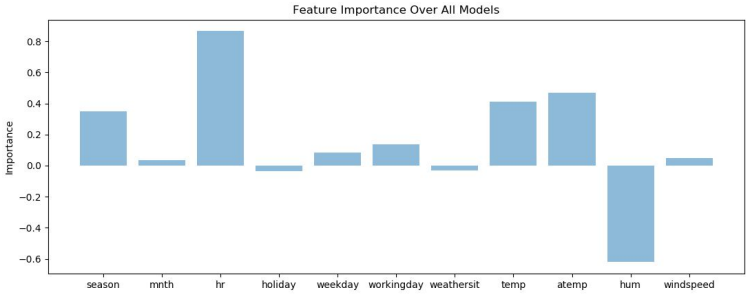
75 Divisions



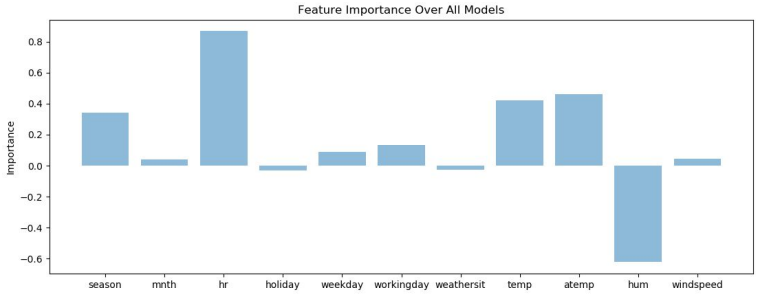
100 Divisions



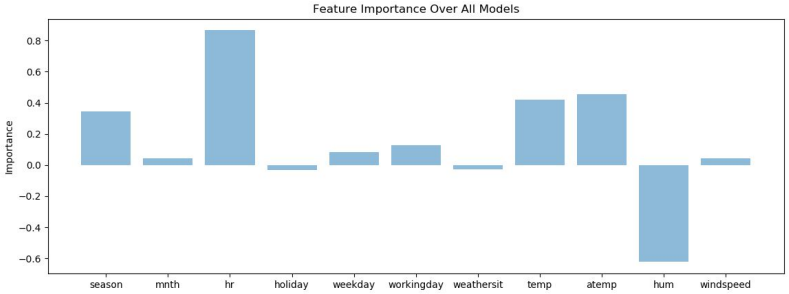
Feature Analysis: Logistic Regression



100 Divisions



250 Divisions



500 Divisions

Summary

- Best:
  - Random Forests and Logistic Regression
- Worst:
  - Linear Regression
- All exhibited expected behavior
- All of the models showed Hours as the most important feature

Conclusion and Future Work

- Confirmed hypothesis
- What about perfect accuracy??

Future:

- Larger Datasets
- Neural Networks
- Harder Regression/Classification Problems

(Discovery Communications LLC)

Questions?

Thank You