Natalie Carlson
CSCE 415
Project 2
March 2019

1.
The assumptions, or inductive bias, in this project included how the data was split, which polynomial degrees were tested, and ultimately the degree selected and applied to the test set. The data was split into thirds. Data point one was assigned to the training set, then data point two to the validation set, and then data point three to the test set, and then the continuing of this pattern until all data points were in one of the three sets. The validation set was modeled against the training set with polynomial regressions of degree 1, 2, 3, 5,10, 15, & 20. The selected function chosen for the test set was degree 3, as the results showed it had the best performance overall in the training against the validation set step while also producing a positive number of forecasted homes.

2. See figure 4.
Forecasted number of homes per year:
2019 2552.137186432723
2020 2486.109908964485
2021 2418.1702620158903
2022 2348.3182455878705
2023 2276.5538596790284
2024 2202.8771042912267
2025 2127.287979424
2026 2049.786485075485
2027 1970.3726212484762
2028 1889.0463879406452

3. See figure 5.
Validation set against training set:
Mean squared error for degree 1 : 7498596.37109375
Variance score for degree 1 : -0.0004372812663073766
Mean squared error for degree 2 : 6408688.185424242
Variance score for degree 2 : 0.14497457025091887
Mean squared error for degree 3 : 5135638.653239128
Variance score for degree 3 : 0.31482051872818717
Mean squared error for degree 5 : 3693911.937522695
Variance score for degree 5 : 0.5071708046244219
Mean squared error for degree 10 : 5726717.602141326
Variance score for degree 10 : 0.23596077119823589
Mean squared error for degree 15 : 9395261.016709026
Variance score for degree 15 : -0.2534838419330341
Mean squared error for degree 20 : 18673739.673722256
Variance score for degree 20 : -1.4913869777376112

Test set against training set:
Mean squared error for degree 3 : 5266975.674070025
Variance score for degree 3 : 0.27624534412413637

4.
Figures 1 and 2 demonstrate how the data selection can impact the model. Although the data was split evenly, there is a large difference in the data values of the validation set compared to the training set and the test set compared to the training set.

Figures 3 and 5 show the importance of developing multiple models to prevent choosing an algorithm that is underfit or overfit.

The results are constricted by not only the limited amount of data, but also by having only a single factor to base future prediction on. Number of homes built per year are affected by mortgage rates, construction rates, average household incomes, the real estate market, unemployment rates, available land, availability and costs of supplies, and much more. Additionally, there are unforeseen events and factors that also impact the actual results. Increasing the number of variables and the number of examples to train against would greatly improve the algorithm.

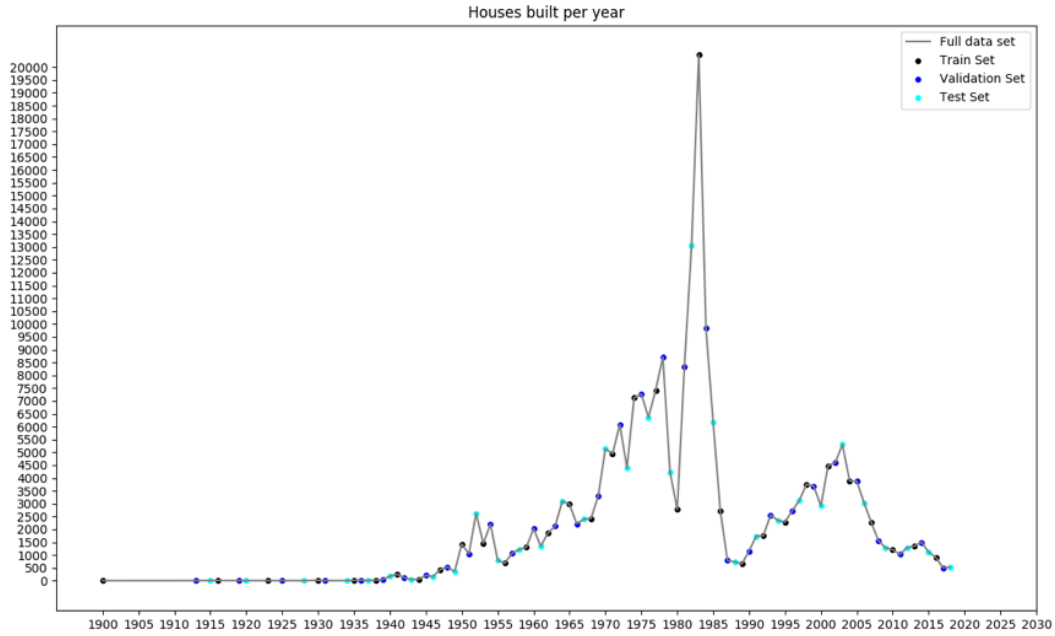

Figure 1

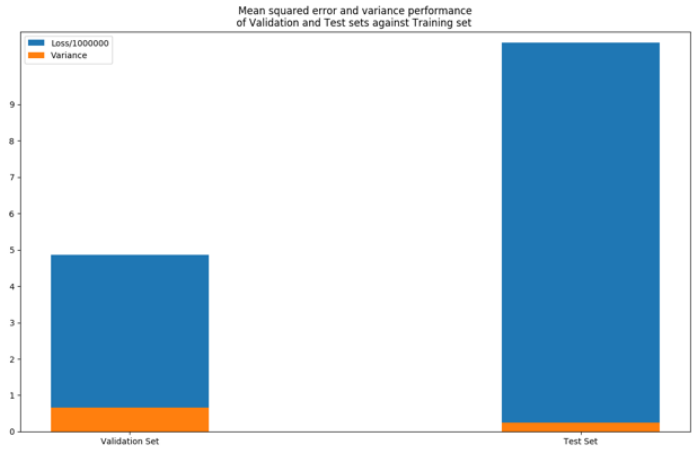

Figure 2

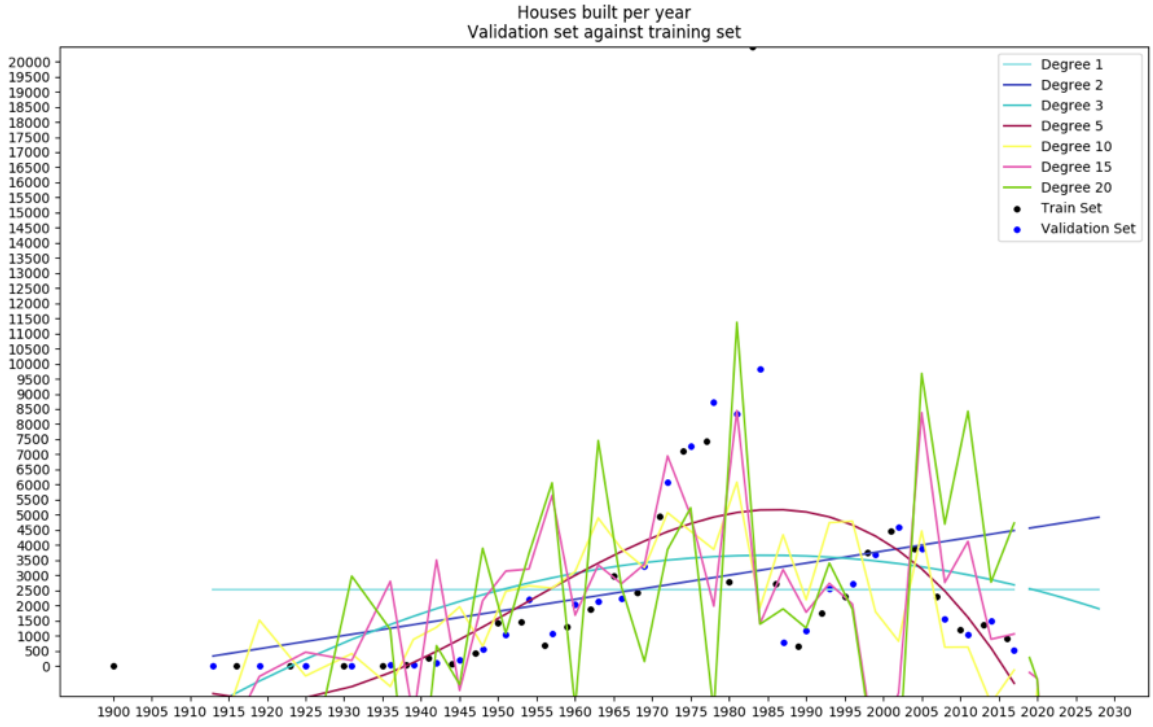

Figure 3

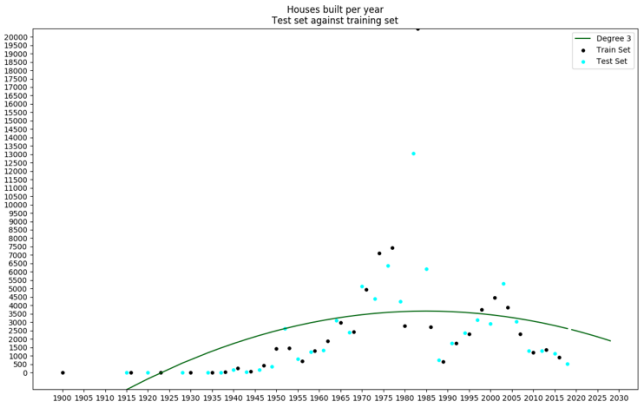

Figure 4

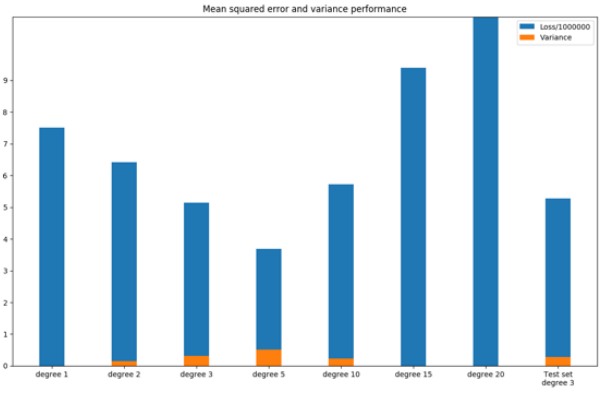

Figure 5