HW2 Question 6:

Write a 2-3 page report about your favorite model. The report should include:

- A description of how it works.
- An evaluation of how well it works.
- Any interesting experiences or surprises you had over the course of these experiments.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm that builds a linear model such that the data points are separated by a vector (or a hyperplane). This vector tries to maximize the distance/margin between the points (or groups) on either side of the vector(s). This approach can be used for both classification and regression (Support Vector Regression) . This algorithm is directly proportional to the square of number of points and linearly to the number of attributes or features.

Given a set of points, the challenge is to identify the correct hyperplane that maximizes the margins.

Step1: Identify all the hyperplanes/vectors
Vector representation of a hyperplane : w.x - b = 0
{w,x,b are vectors}

w -> is the perpendicular vector to the hyperplane
b -> intercept on the axis
w and b can have different set of values that uniquely divide the points into groups.

Step 2: Maximize the margins
The maximum margin is obtained when the hyperplane lies exactly halfway between the group of points.
w.x - b = 1 -> hyperplane at extreme 1
w.x - b = -1 -> hyperplane at extreme 2
Midway between the above two will be w.x - b = 0

The distance between these two hyperplanes is 1/||w||
||w|| -> scalar value of the vector
To maximize 1/||w||, we need to minimize ||w||.

Step 3: For each point added to the plane, find midway (optimal ||w||) by computing the cost and update the hyperplane.
Repeat step 3 till all the points are processed

For SVM Regression, a threshold (epsilon) is used to monitor and penalize errors with the loss/hinge function (used to maximize the margins) . Epsilon defines a range within which no penalty is associated in the training hinge function with points predicted within a distance epsilon from the actual value.

SVM performance on Zillow Dataset

1. SVM performed almost as good as the other models like Linear Regression, K Nearest Neighbor (Highest variance) . and better than Random Forest (got negative correlation for RF)
2. The SVM built was a linear model and without kernels
3. Mean squared error: 0.02881
   Variance : 0.00285
4. The parameters, epsilon (specific to SVM regression) was modified to lower value to avoid overfitting of the data.
5. SVM Radial Basis Function kernel needs to be explored but it might be prone to overfitting in the kernel.

Experience with Zillow Challenge and modeling

1. A lot of features can be eliminated by just going through its description in the data_dictionary. Around 15 features carried redundant information.
2. There were more attributes that had no information or insufficient information to extrapolate it to other missing values.
3. Since the dataset was of reasonable size, we could afford to drop rows that had NaNs.
4. Z-Normalizing the dataset helped the algorithms to come up with better models than when done with unnormalized dataset.
5. The models were easy and fast to train on the machines
6. Tree based regressions didn’t perform well (Random Forest produced negative correlation)
7. Calculatedfinishedsquarefeet attribute played the most crucial role in the model
8. SVM was the most computing intensive algorithm amongst the Linear Regression, K Nearest Neighbor.
9. The lower the epsilon value the more time it took for the model to be trained.
10. The correlation of epsilon and the computing time is likely due to the low margin for loss function that’s harder to find.