**Part II. Programming**

**The first part is data preprocessing**:

After replacing missing values, **line 63** starts to separate data into training, validation and test set based on requirements.

A close up of a map

Description generated with high confidence

**The second part is one variable least squares model**:

**Line 87** starts to build model based on equation one variable OLS model, which is the function named as best fit slope and intercept. Then function names as rmse is applied to calculate the Root Mean Squared Error(RMSE). The equation for the line is:

y = 112.414\*x+12662.054

**The third part is all features least squares mode**l:

**Line 119** starts to build all features least squares model based on imported package, which is from sklearn.linear\_model to import LinearRegression.

Before training the model, we are required to transform the categorical features to one-hot encoding so that they can be used in the model. **Line 123**, we employ panda get\_dummies to transform categorical features. We should pay attention to this case that we only need k-1 dummy variables to indicate a feature with k variables. Hence, drop\_first=True is applied. (finally, there are 330 feature columns)

Then we train all features least squares model with the training set, validate it with the validation set.

**The forth part is Lasso regression model**:

**Line 151** starts to build Lasso regression model based on imported package, which is from sklearn.linear\_model to import Lasso.

Before training the model, we are required to normalize training, validation and test set by substracting training set mean and dividing by training set std. **Line 153** starts to normalize those data sets.

**Line 164** starts to train Lasso regression model. To employ cross validation, we concatenate training and validation set, then separate them into 4 folders. For each Alpha, we train the model by randomly picking three of folders and validate it by the rest one. Then, we average the RMSE calculated by these 4 folders.

After training and validating, the RMSE for each alpha is listed. And we could find that the lowest error on the validation set appear at Alpha=400.

As graph shown, the RMSE for training data keeps decreasing, however, the RMSE for validation data simultaneously decrease before Alpha=400 and starts to increase after Alpha=400. This is the beginning of over-fitting. The phenomenon named as over-fitting is that the model exactly fits the training set but fails to generalize to validation and test sets.

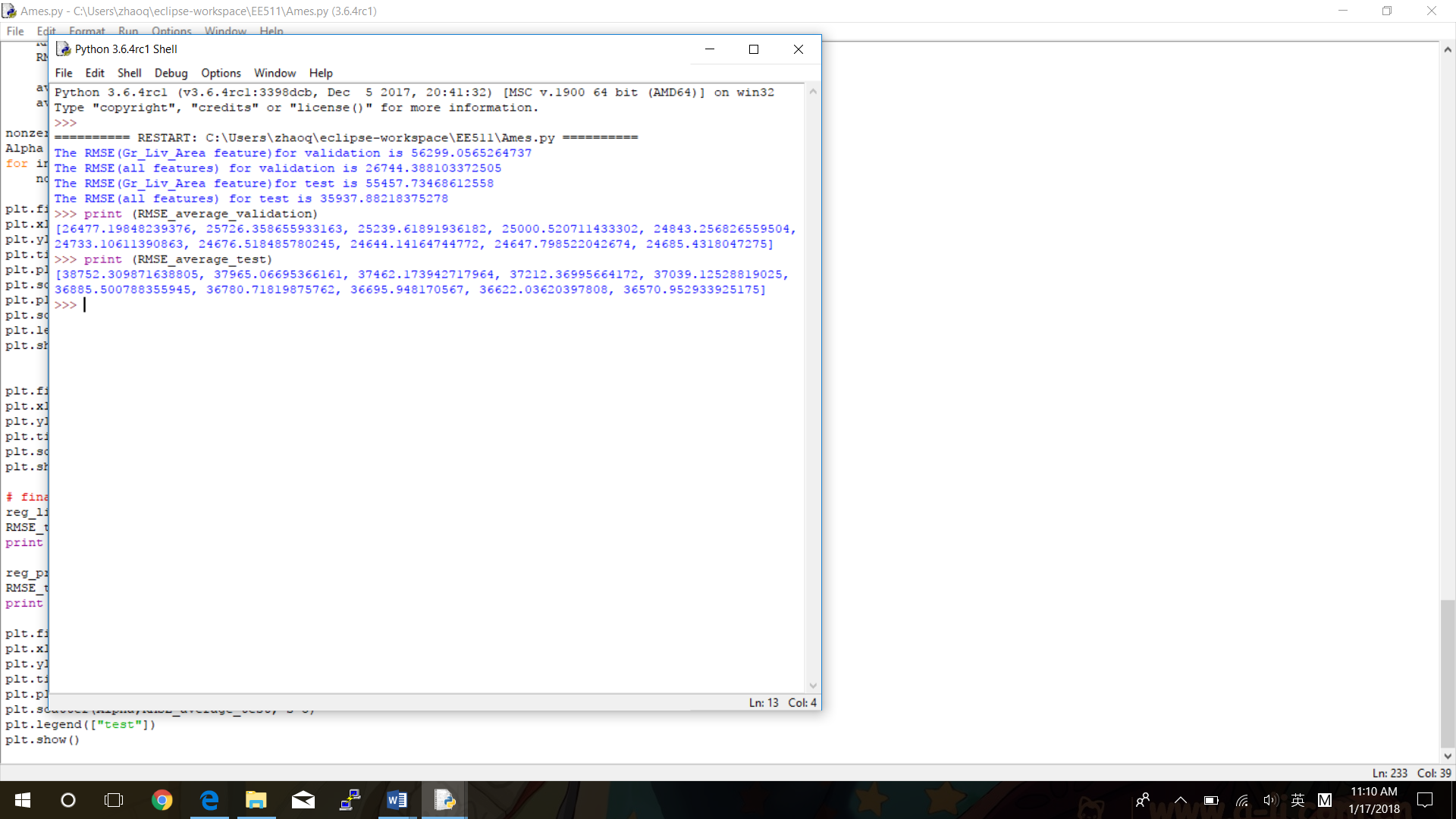
A close up of a map

Description generated with high confidenceA screenshot of a cell phone

Description generated with very high confidence

**The fifth part is using the test set:**

Line 226 applies the single variable model, the all features least squares model and the regularized model to the test set. The RMSE for each condition for both the validation set and the test set are listed in the table.



|  |  |  |
| --- | --- | --- |
| Model\RMSE | Validation Set | Test Set |
| Single variable model | 56299.056 | 55457.734 |
| All features Least Squares Model | 26744.388 | 35937.882 |
| Lasso Regression Model (Alpha =50) | 26477.198 | 38752.310 |
| Lasso Regression Model (Alpha =100) | 25726.359 | 37965.067 |
| Lasso Regression Model (Alpha =150) | 25239.619 | 37462.174 |
| Lasso Regression Model (Alpha =200) | 25000.521 | 37212.370 |
| Lasso Regression Model (Alpha =250) | 24843.257 | 37039.125 |
| Lasso Regression Model (Alpha =300) | 24733.106 | 36885.501 |
| Lasso Regression Model (Alpha =350) | 24676.518 | 36780.718 |
| Lasso Regression Model (Alpha =400) | 24644.142 | 36695.948 |
| Lasso Regression Model (Alpha =450) | 24647.799 | 36622.036 |
| Lasso Regression Model (Alpha =500) | 24685.432 | 36570.953 |

The RMSE for validation data decreases before Alpha=400 and starts to increase after Alpha=400. Alpha=400 is the beginning of over-fitting. The phenomenon named as over-fitting is that the model exactly fits the training set but fails to generalize to validation and test sets.