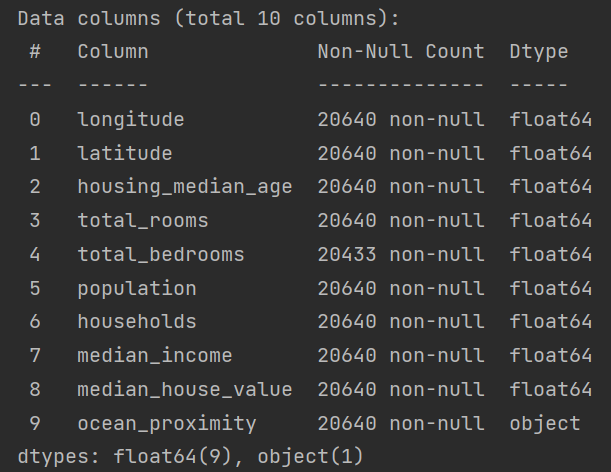
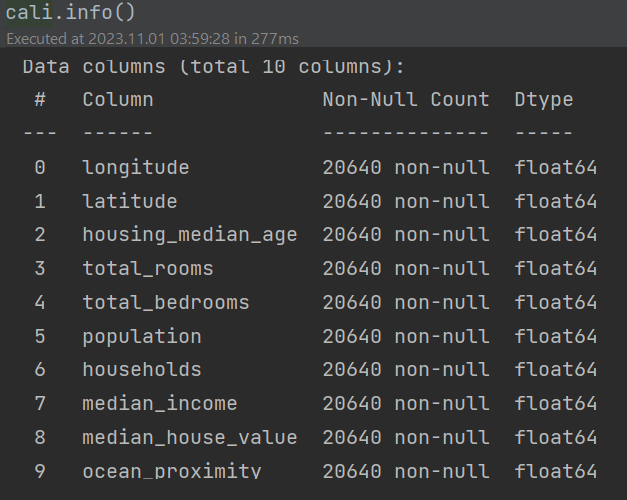
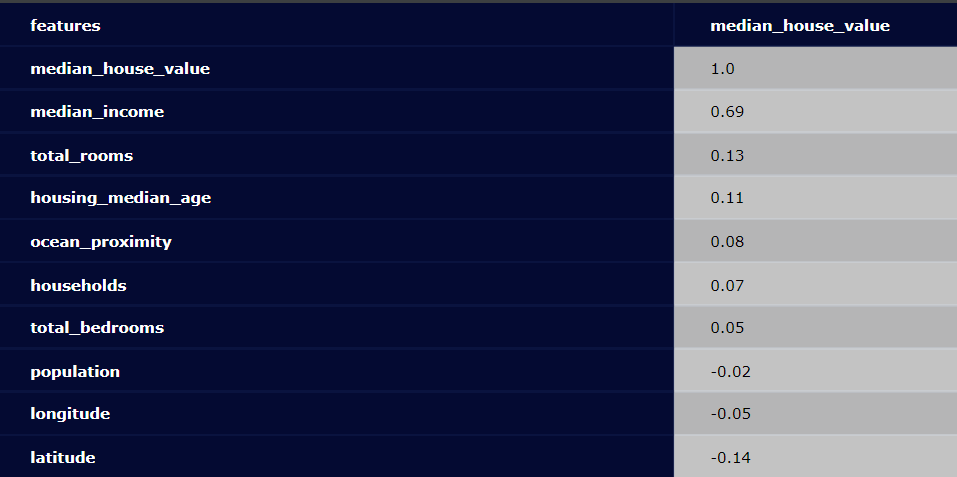
For building a regression model to predict the price of a house in California, I have followed the next steps.

* Cleaning and Normalizing Dataset



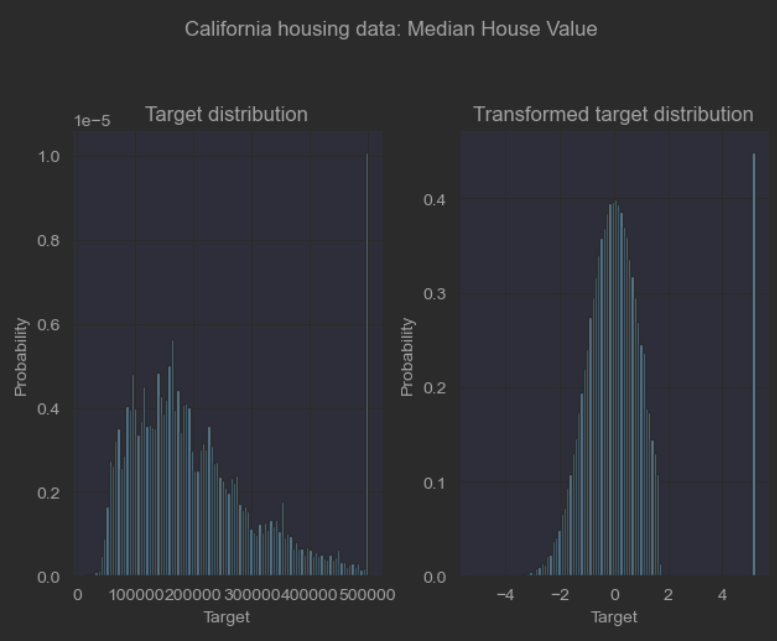


* + OrdinalEncoder has been used for normalization of ‘ocean\_proximity’ feature
  + For NaN values has been used median mode for fillna method
* Looking for corelation

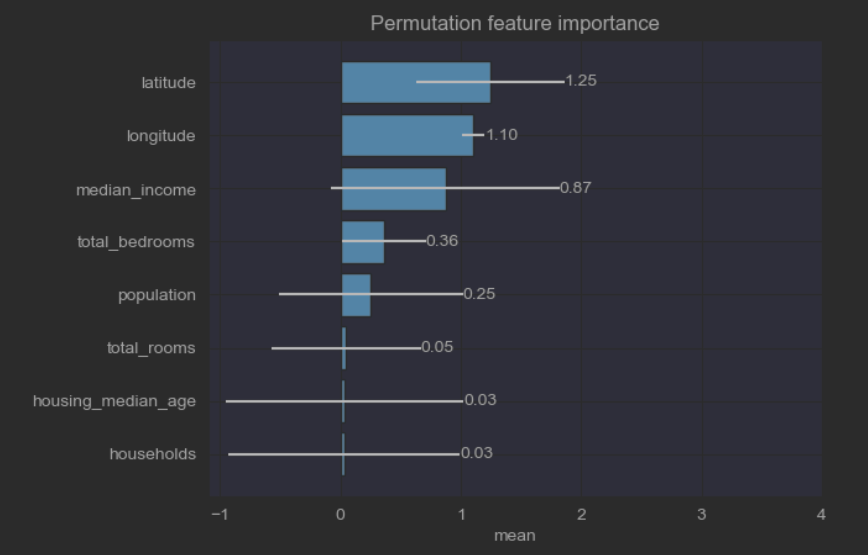


Observations:

* + median income seems to be an important factor with a correlation of about 0.7
  + next biggest but small correlation of about 0.1/-0.1 with the rest features
  + I think the lower negative latitude correlation can be explained with that lower latitude values indicate more closeness to the coast, and the coast tends to have higher housing prices.
* normalize the target distribution



* + QuantileTransformer is used to normalize the target distribution before applying a RidgeCV (LinearRegression) model, the result was worst, not used in the model.
* Permutation feature importance



From the plot we can see latitude, longitude and median\_income are of importance.

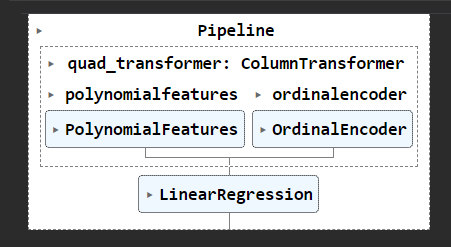
* Optimal Model Complexity



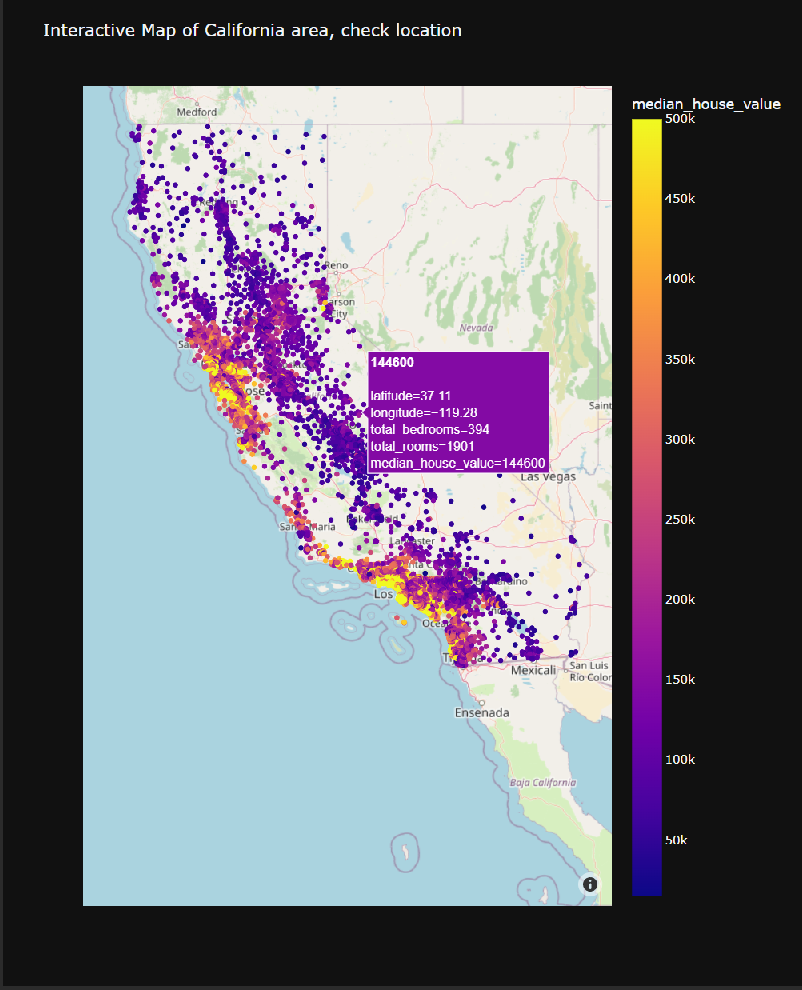
The best degree of complexity is 2, which is not very clear from the above plot



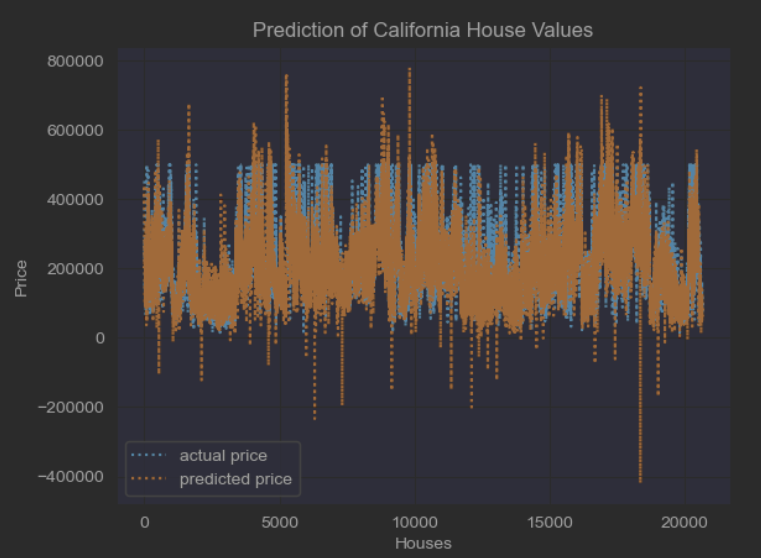
* + The model has been trained with five sequential Polynomial degrees, looking at the mean-square-error(test\_mse) the smallest test error is for degree 2. So we will will build a quad Polynomial model
* Quad Polynomial Model



* **Conclusion** 
  + The model performed better comparing test\_mse = 4.371574e+09 against the mse\_baseline\_test = 1.332592e+10, also the training set has better characteristics against the baseline training set; train\_mse=4.371574e+09 and mase\_baseline\_train=1.331026e+10.
  + The R-sq is 0.704614, which indicates that the regression predictions perfectly fit the data.
  + Train/Test split is test\_size=0.33, random\_state=42
* Plots of Predict, Model and Actual House Prices
  + Actual

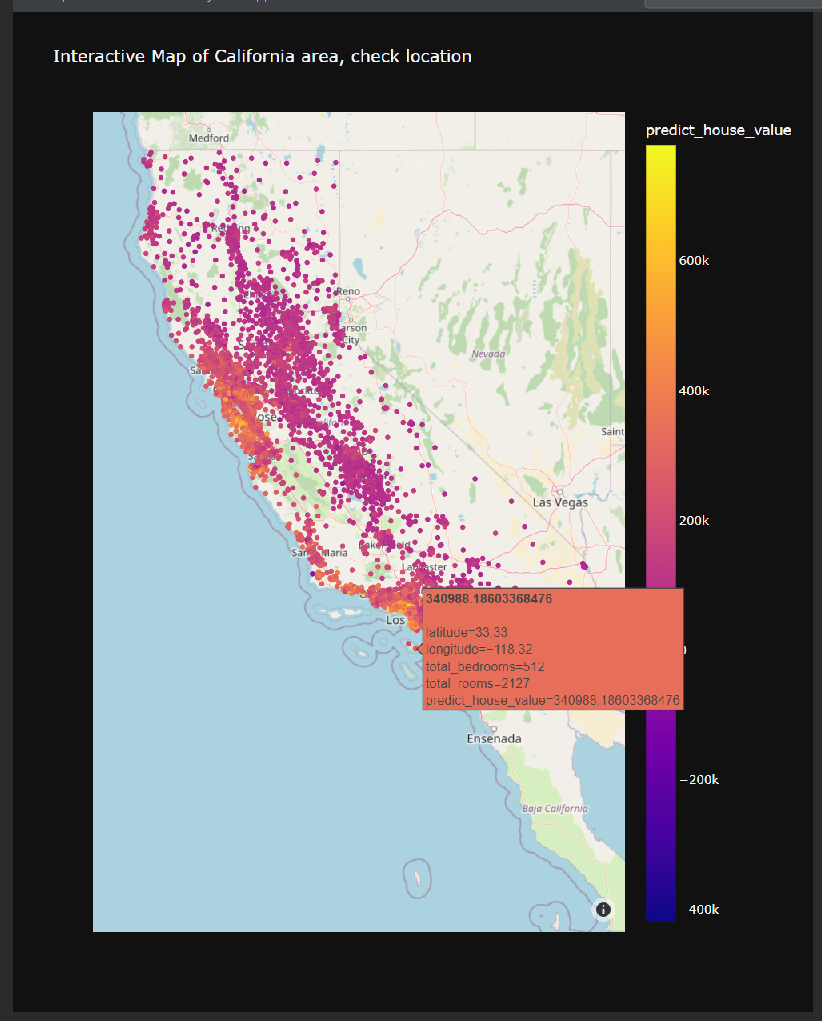


* + Predict

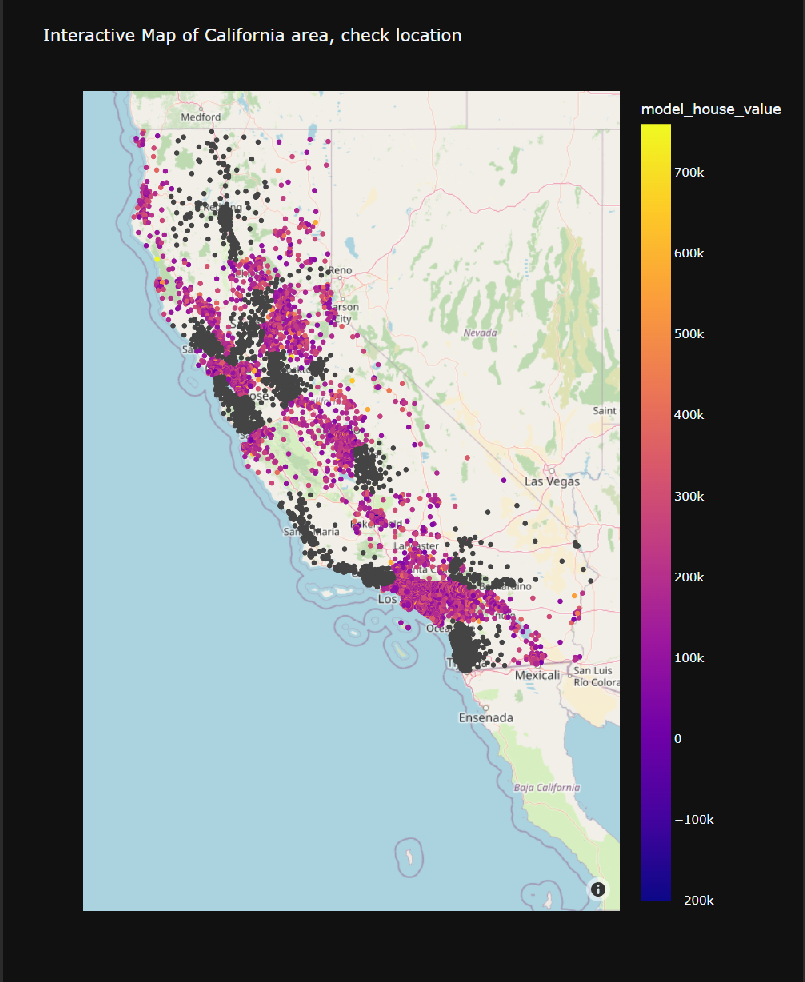


From the above plot we can see the actual house value are capped, also nice a prediction for house values, probably influenced by the permutation features importance as latitude, longilude and median\_income.

We see plenty outliers, it will be nice for which area the model predicts that negative house price, it needs further investigation.



* + Model



As the the data set is split  on train and test sets, the abobe plot explains at 67% of data, so we can see which area have not been covered by the model.