

This project uses the California Housing Prices dataset at <https://www.kaggle.com/camnugent/california-housing-prices>. The data contains information from the 1990 California census. It has an easily understandable list of variables, and has 20640 rows of data, which is manageable in size.

I have borrowed heavily from Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'.

## Data Cleaning and Wrangling

There are 207 non-available values in the total-bedrooms column, and the correlation of total-bedrooms to medium\_housing\_value is very low, at 0.049686. Therefore, I have decided to drop total-bedrooms from the dataset.

The ocean\_proximity column contains only string values. To make it easier to plot, I have replaced the string values by integers['ISLAND'=0, 'NEAR BAY'=1, 'NEAR OCEAN'=2, 'H OCEAN'=3, 'INLAND'=4].

Since the correlations of medium\_housing\_value to longitude(-0.0459), population(-0.0246), and households(0.065843) are also low, I have dropped these 3 columns as well.

The median\_income column is measured in tens of thousands of US Dollars. I have converted it to US Dollars.

## Different Models

Using Statsmodel to fit Linear Regression is very easy and straightforward. Sklearn, on the other hand, produces somewhat close results. The f-statistics are about 4586 and 4341, respectively.

The Decision Tree Regression is not suitable for this dataset, the f-statistic is 2776.

The Random Forest Regression, on the other hand, produces the best result. The f-statistic is 7650.

The SVR has f-statistic 1373.

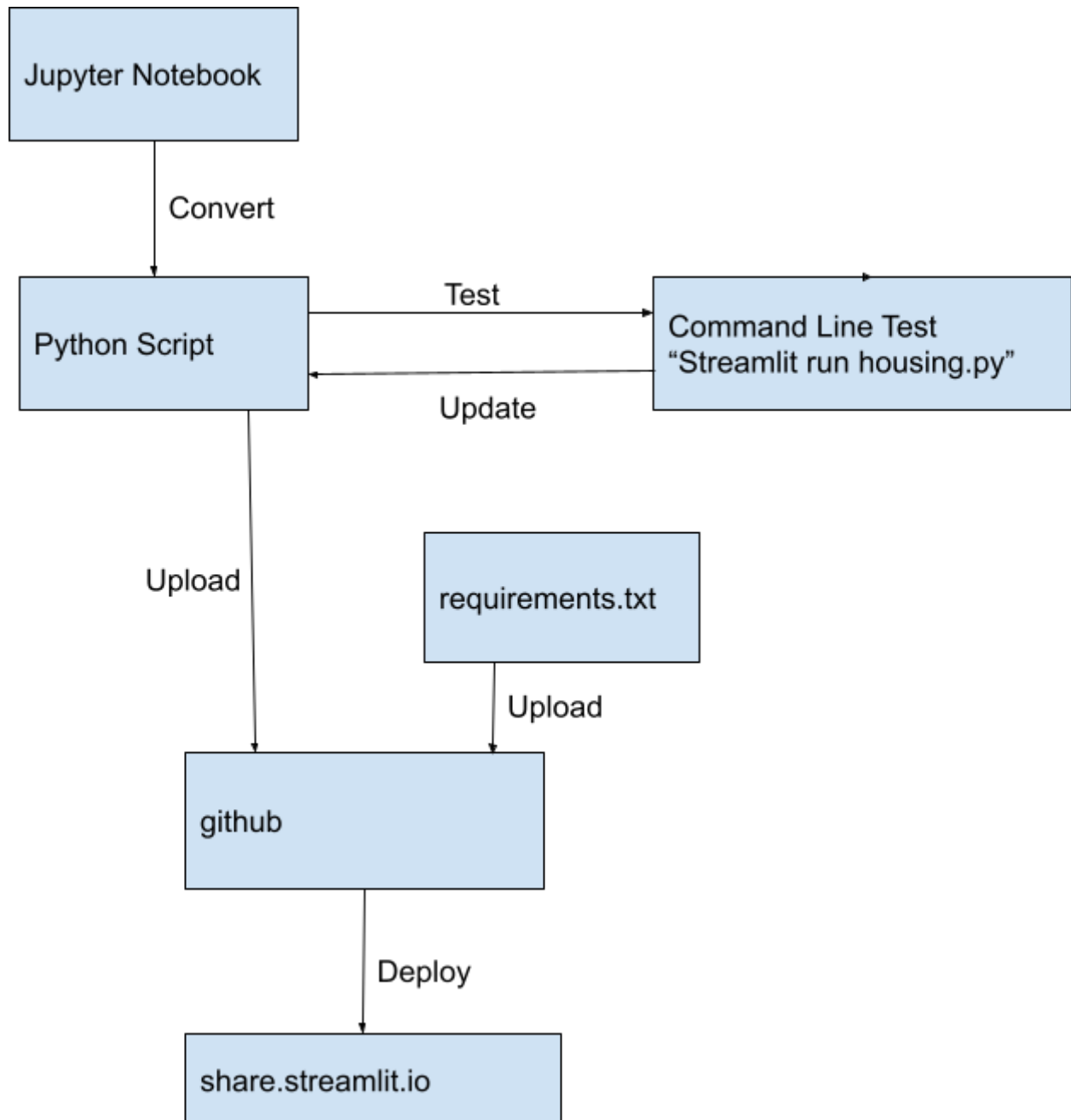
## Model Fine Tuning

I have used GridSearchCV and RandomizedSearchCV to fine-tune the Random Forest Regression model. The f-statistics has reached 7723 and 7964, respectively, which are better than 7650 as before.

## Deployment

I have decided to use streamlit to deploy my project, because it is the easiest in the market. I am aware that there are far more mature tools out there for machine learning deployment, I would like to investigate more once I have time.

Here is the diagram to illustrate the deployment steps:



## Project Files

### Code URL

<https://github.com/tiffany88888/capstone.git>

|   |   |
|---|---|
| Data File                                       | housing.csv.zip   |
| Jupyter Notebook File                           | capstone.ipynb  |
| Jupyter Notebook file (using postgres database) | capstone_scaled2.ipynb  |
| Python Script                                   | housing.py  |
| Requirements File                               | Requirements.txt  |
| Readme File                                     | README.pdf  |
| Introduce the Web Tool GUI                      | introduction_post.pdf   |
| Screenshots                                     | <ul style="list-style-type: none"><li>• Run_streamlist.png</li><li>• California_housing_price.png</li><li>• California_housing_price_2.png</li><li>• California_housing_price_3.png</li></ul> |

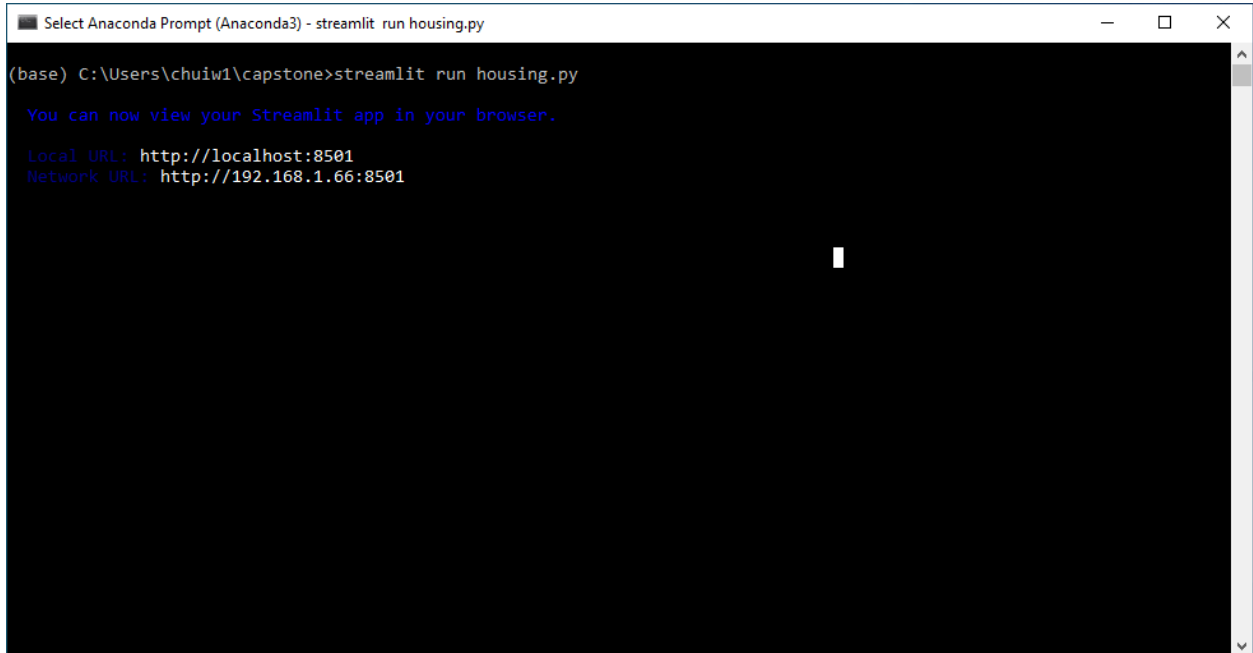
### Deployed URL

<https://share.streamlit.io/tiffany88888/capstone/main/housing.py>

## Screen Shots

To run as Web App Locally (Run\_streamlist.png)

- Use command “streamlit run housing.py”



```
Select Anaconda Prompt (Anaconda3) - streamlit run housing.py

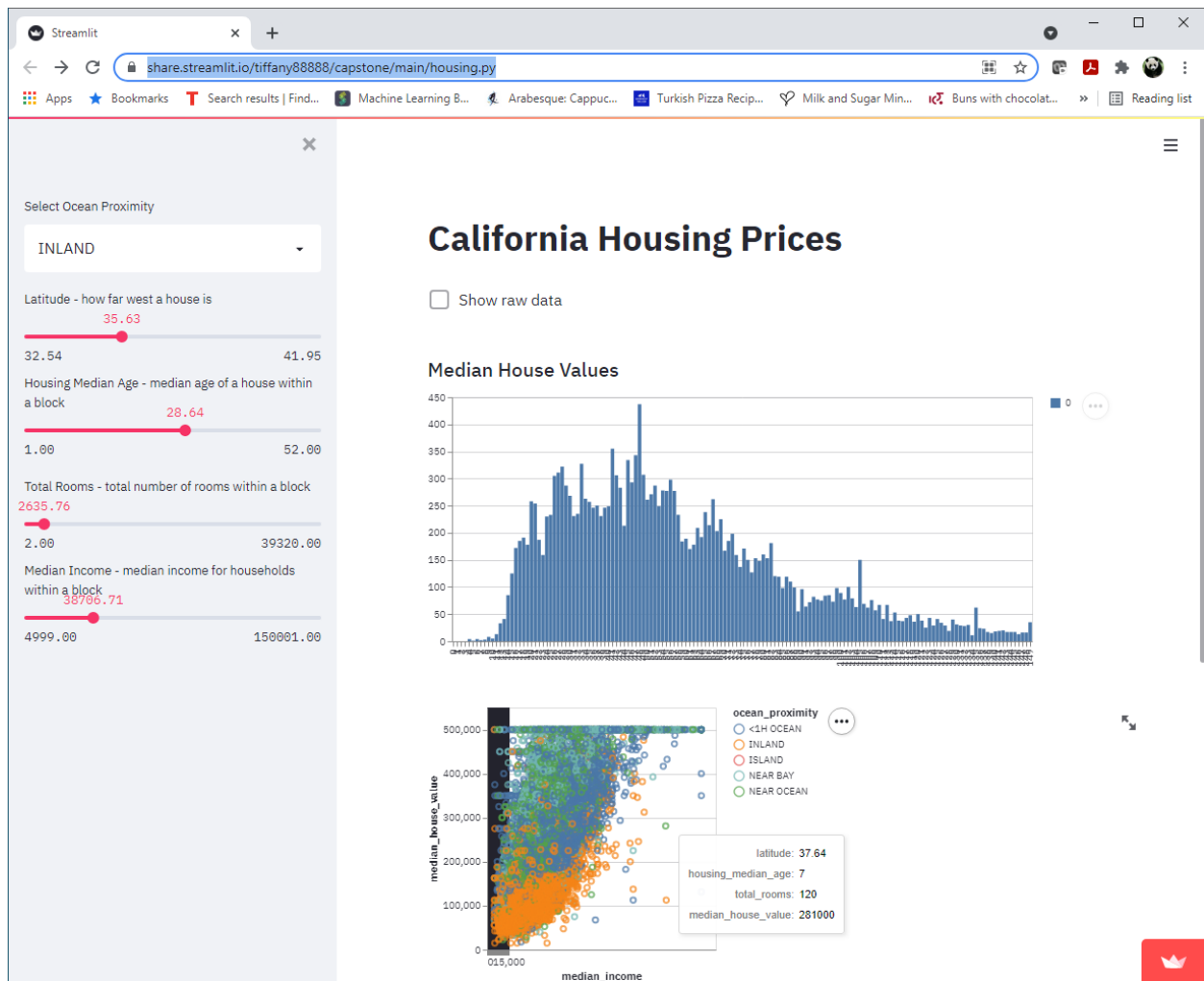
(base) C:\Users\chuiw1\capstone>streamlit run housing.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.1.66:8501
```

Web App - Page One (California\_housing\_price.png)

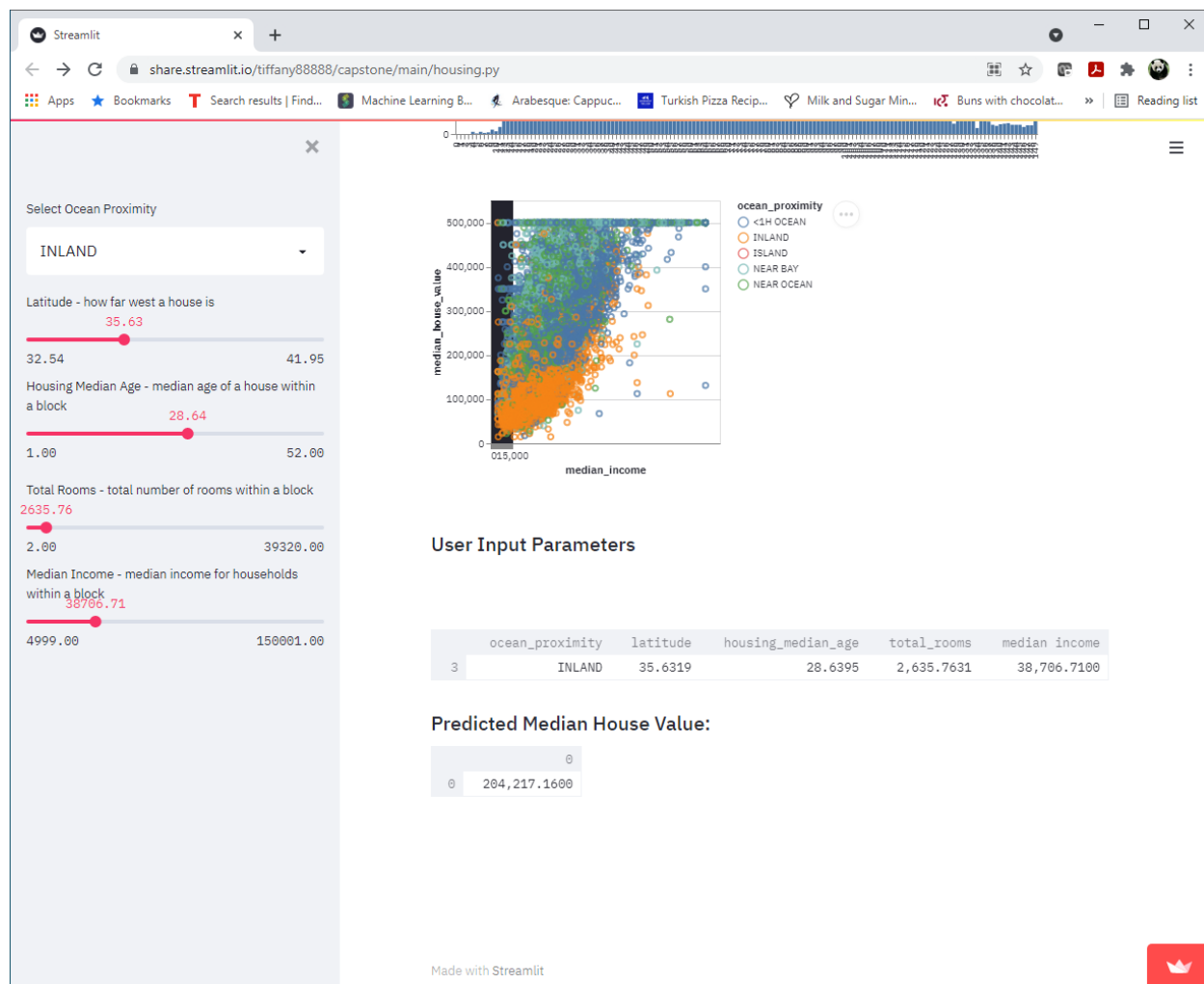
- User Input Panel on the left
- Median\_House\_Value Histogram on top
- Median\_Income vs Median\_House\_Value Chart at the bottom



## Web App - Page Two (California\_housing\_price\_2.png)

- User Input Panel on the left
- Median\_House\_Value Histogram on top
- Median\_Income vs Median\_House\_Value Chart in the middle
- User Input Parameter Summary and Predicted Median House Value at the bottom

Users may change the parameters on the left to obtain different predictions.



Streamlit

share.streamlit.io/tiffany8888/capstone/main/housing.py

Apps Bookmarks Search results | Find... Machine Learning B... Arabesque: Cappuc... Turkish Pizza Recip... Milk and Sugar Min... Buns with chocolat... Reading list

SELECTED: INLAND

Latitude - how far west a house is

35.63

32.54 41.95

Housing Median Age - median age of a house within a block

28.64

1.00 52.00

Total Rooms - total number of rooms within a block

2635.76

2.00 39320.00

Median Income - median income for households within a block

38768.71

4999.00 150001.00

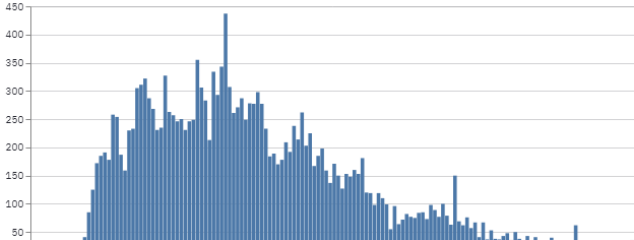
## California Housing Prices

☒ Show raw data

### Raw Data

|    | latitude | housing_median_age | total_rooms | median_income | median_house_va |
|----|----------|--------------------|-------------|---------------|-----------------|
| 0  | 37.8800  | 41                 | 880         | 83252         | 45              |
| 1  | 37.8600  | 21                 | 7099        | 83,014.0000   | 35              |
| 2  | 37.8500  | 52                 | 1467        | 72574         | 35              |
| 3  | 37.8500  | 52                 | 1274        | 56,431.0000   | 34              |
| 4  | 37.8500  | 52                 | 1627        | 38462         | 34              |
| 5  | 37.8500  | 52                 | 919         | 40,368.0000   | 26              |
| 6  | 37.8400  | 52                 | 2535        | 36591         | 29              |
| 7  | 37.8400  | 52                 | 3104        | 31200         | 24              |
| 8  | 37.8400  | 42                 | 2555        | 20804         | 22              |
| 9  | 37.8400  | 52                 | 3549        | 36912         | 26              |
| 10 | 37.8500  | 52                 | 2202        | 32031         | 28              |

### Median House Values



A histogram showing the distribution of median house values. The x-axis represents the median house value, ranging from 0 to 150,000. The y-axis represents the frequency, ranging from 0 to 450. The distribution is unimodal and slightly right-skewed, with a peak frequency of approximately 440 occurring at a median house value of about 35,000. The data is represented by blue vertical bars.