

**CS4055  
Data Mining and Data Warehousing  
Project Report**

***California Housing Prices Prediction***

Aidan Cleere 15160009  
Colm Le Gear 15148823  
William O’Leary 15155528  
Aaron Dunne 15148602

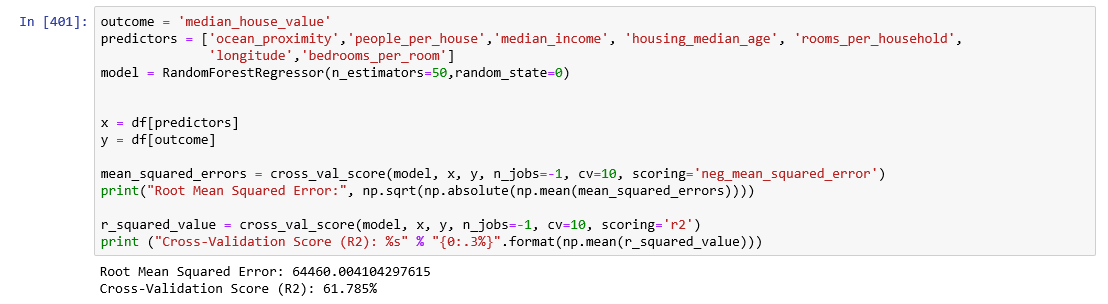
**Contribution**

|  |  |
| --- | --- |
| Aidan Cleere | 25% |
| Colm Le Gear | 25% |
| William O’Leary | 25% |
| Aaron Dunne | 25% |

# **Introduction**

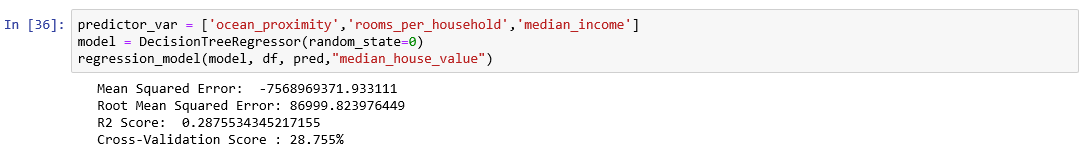
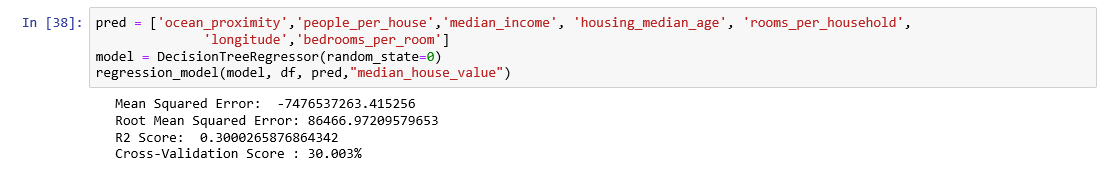
For our dataset, we chose California Housing Data (1990) from Kaggle the dataset has 20000 rows and 10 columns (Sharma, 2018). Our aim was to predict the median house value. The dataset contains details on housing blocks, containing information such as the number of people who lived in that area, the total bedrooms in the area and the median income of the residents.

# **Predictive Models**

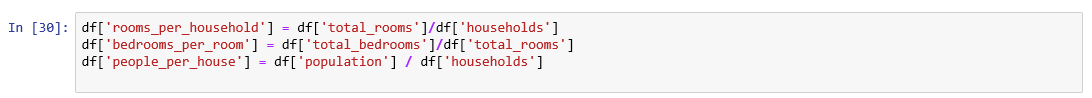
Our final model was a Random Forest. This proved to be our best performing model, giving us an R squared (Cross-Validation) score of 61.785% (Fig. 1).

**Fig. 1** – Random Forest Model

Our other models were Linear Regression (Fig. 2a) and Decision Tree (Fig. 2b), both of which didn’t achieve a cross-validation score any higher than 38%. We didn’t achieve a very high accuracy when predicting our target variable, but we believe this is due to the data not providing enough info needed for an accurate prediction.

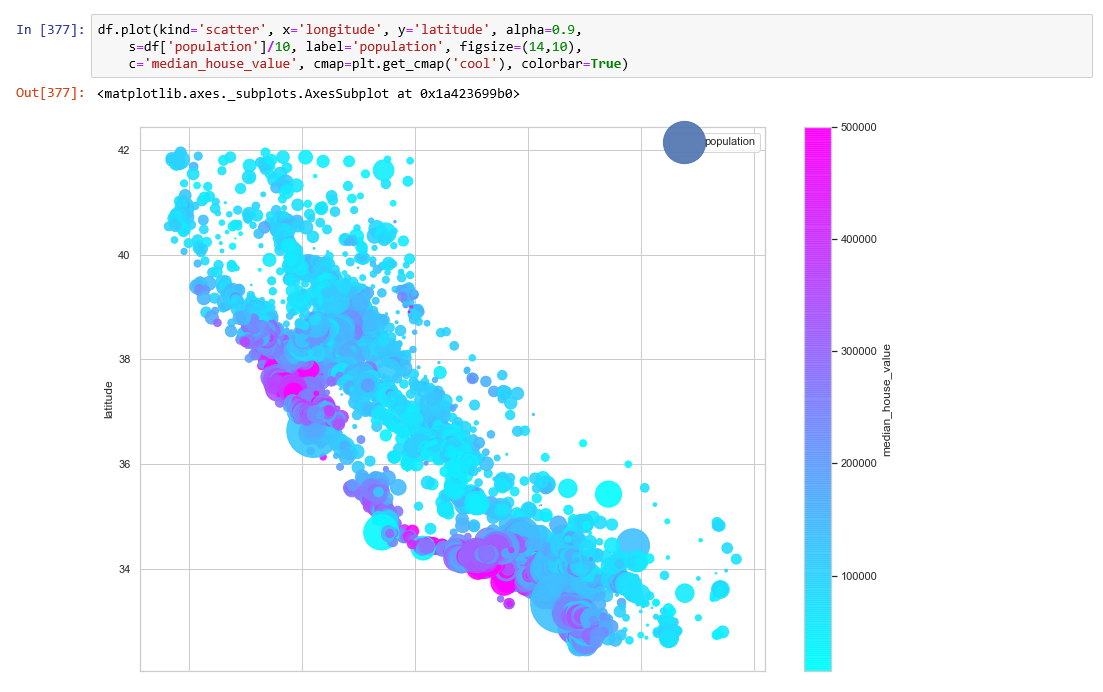
**Fig. 2a** - Liner Regression Model

**Fig. 2a** – Decision Tree Model

To improve our data, Colm created 3 new features: rooms per household, bedrooms per room and people per house (Fig. 3). These new features made a decent improvement upon our accuracy score.

**Fig 3**- Newly created features to improve on results.

We also removed any data which was on an island as all the island data were outliers and skewed our data. When we fit our models, we scaled our data to negate the effect of the large variance in our prediction values.

When we compared the heatmap of median house values plotted using longitude and latitude, we found that the higher values were to the left of the graph. When you look at the graph where the population is plotted over a scatter plot (Fig. 5), we see that the left of the graph is the coast. This indicated that the median house value increased the closer to the coast you went. Some quick analysis of the data showed this to be true.

**Fig. 5** – Scatter plot of the population across California

**References**

Sharma, V., 2018. *California Housing Prices Prediction System (ML).* [Online]   
Available at: https://www.kaggle.com/vikassharma807/california-housing-prices-prediction-system-ml/notebook  
[Accessed 30 November 2018].