Data Science for Innovation

Assessment Stage 2 | Report

Agustin Ferrari | Nathan Collins | Yasaman Mohammadi | Luca Sardo

Assessment 3:	Literature review related to your approach Defining an experiment setup Implementing your proposed approach Evaluating and analysing your approach Conclusion and any future work
Type:	Group Assessment
Length:	3000 words
Weight:	30%
Due:	May 19 at 23:59

Section 1: Literature Review

[1.1] **Ye, T., Johnson, R., Fu, S., Copeny, J., Donnelly, B., Freeman, A., Lima, M., Walsh, J., & Ghani, R.** (2019). *Using Machine Learning to Help Vulnerable Tenants in New York City.*

"Ye, T." *et al*. explore their collaborative effort with the New York City Public Engagement Unit (NYC PEU) following machine learning applications in identifying tenants subjected to landlord harassment. The finalised model could ascertain **59% more** buildings where tenants face landlord harassment than current outreach methods with the same resources. The investigation further illustrates a comprehensive, data-driven methodology, one that facilitates an improved outreach to vulnerable tenants while predicting tenant-harassment risk factors.

> Application of Machine Learning
>
> *Scikit-learn*'s **Random Forest** and **Decision Tree** libraries were first implemented to determine feature importance. The authors subsequently plotted the top 20 significant features with a **gradient-boosting** model, ascertaining primary predictors within high-risk building complexes.

The interpretability of the feature importance by "Ye, T." *et al*. is considered significant, as it may be applied to determine driving mechanisms of rent stress, offering insights into the relationships between input features and their target variables. By selecting a highly interpretable model, further meaningful material about primary predictors of rent stress can be extrapolated. While gradient-boosting models outperform current outreach practice, ensemble models may be necessary for training a robust predictor.

[1.2] **Muthukrishnan, R., & Rohini, R.** (2016, October). *LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 18-20). IEEE.

The authors examine regression analysis, a widespread machine learning technique applied in parameter estimation and variable selection. The paper explores conventional feature selection methods, concentrating on **LASSO**, **Ridge** and **OLS** regression. Experimental findings in this paper reveal that LASSO typically

outperforms Ridge and OLS through coefficient reduction (to zero), though it is proposed only as an alternative to conventional feature selection methods.
While the article doesn't study rental stress or rental vulnerability outcomes, the authors empirically elucidate a fundamental and optimised regression approach to integrate within the preceding investigation. It is anticipated that the variable selection component will petition optimisation, as the census dataset's complexity and multi-faceted description render it paramount to first clarify the relevant variables that proliferate rental stress. As such, integration of the LASSO approach would be applied to yield the project's key variables.

[1.3] **Keith Jacobs, Rowland Atkinson, Val Colic Peisker, Mike Berry and Tony Dalton. (2010, September).** *What future for public housing? A critical analysis.* **AHURI. Report. 151.**

"Jacobs, K." *et al*. investigate a decline in support for Australia's public housing system. Their claims are illustrated by a political influence, asserting that existing schemes are difficult to reform and remain unattractive from an investment perspective. As tenants benefitting from these schemes typically affiliate with marginalised communities, these represent small population percentages with poor political influence and agency; improperly representing its impact.

The adverse reception of public housing spending from the wider population creates additional barriers to scheme improvement and revision; as middle-class homeowners typically express resistance to proximal housing developments due to the nature of occupants and diminishing property evaluations. Such an unfavourable image associated with public housing stems further from conflicting narratives; one portrays its occupants as individuals lacking prospects and disturbance-prone, and the other is a suspicion that suggests that occupants exploit public resources.

This analysis provides beneficial insights into the nature of groups receiving public housing, which includes correlative features between public housing and rent stress, such as the nature of tenants. Furthermore, the article illuminates differences in perceptions between public and community housing, which offers correlative applicability with the project's rent stress assessment goals.

[1.4] **Tan, J.** (2020). Using Machine Learning to Identify Populations at High Risk for Eviction as an Indicator of Homelessness. (Master's thesis). Massachusetts Institute of Technology.

Tan, J. explores the application of machine learning in identifying eviction-prone populations as a homelessness metric. The article examines features mirroring the current project undertaken, such as age, gender, race, marital history, education, occupation, household income, rent and poverty rate at a statistical area level. It further incorporates the outcome and evaluation of multiple models, namely **Random Forest** and **Linear Regression**.

The Random Forest algorithm is applied in classification and regression tasks, combining predictions of multiple decision trees to improve accuracy and reliability. The article illustrates Random Forest employing a “bagging” technique to create subsets of the dataset and train decision trees on each subset. Random feature selection is introduced to add diversity by randomly selecting a subset of features at each tree node. The final prediction is determined through voting or averaging the individual tree predictions. Random Forest offers advantages such as high accuracy, robustness against overfitting, and feature importance measurement.

Of the models considered in the paper, Random Forest is indicated as the highest performing over Linear Regression. Random Forests furthermore offer a similar attribute, "feature importance", which displays the importance of a feature associated with the outcome variable.

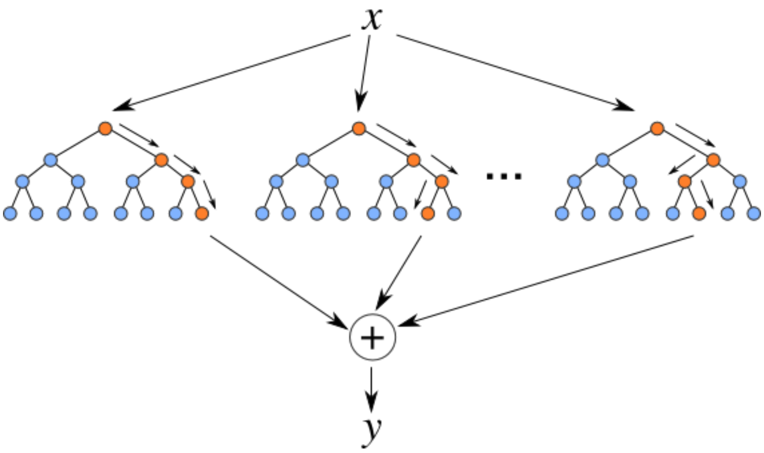


Figure 1 Diagrammatic representation of a Random Forest Regression.

[1.5] **Muir, K., Moran, M., Michaux, F., Findlay, S., Meltzer, A., Mason, C., ... & Heaney, R. A. (**2017). The opportunities, risks and possibilities of social impact investment for housing and homelessness.

The article investigates the dividing features in community housing and public housing, and conducting individual hypothesis tests would be relevant and helpful in identifying rent stress in NSW. This paper provided valuable insight into the characteristics and dynamics of these housing sectors.

Public Housing

Several factors contribute to the sustainability of public housing. Inequality between operating costs and rental income contributes to this problem. According to the Council on Federal Financial Relations, the operating costs of public housing often exceed the income generated by rent. As a result, it is difficult to finance the development of new public housing initiatives.

Furthermore, the ageing housing stock exacerbates the sustainability issue. As housing infrastructure ages, maintenance costs increase, placing additional strain on public housing resources (Council on Federal Financial Relations, 2016a).

Community Housing

There are certain advantages to community housing providers (CHPs), huge ones, compared to state-owned public housing. They can offer a broader range of housing options and support services, access private finance, benefit from tax advantages as charities, and have higher asset management capabilities. Despite this, investment in CHPs has been cautious because of the division of responsibilities between the state and Commonwealth governments. It is important to note that although funding and policies are primarily the state's responsibility, Commonwealth income support payments play a significant role in the revenue base of CHPs.

Section 2: Setup

[2.1a] Research Problem

Research Questions formulated:

*"What are the **variables** or **characteristics** that contribute to rent stress in New South Wales?"*

*"Is there an association between **government interventions** and a **reduction** of rent stress in New South Wales?"*

[2.1b] Hypothesis Testing

As the project's second research question targets the influence of government intervention, the null hypothesis would be established as follows:

Null Hypothesis 1: There is no significant relationship between the predictor (community housing provider) and rent stress.

Alternative Hypothesis 1: There is a significant relationship between the predictor (community housing provider) and rent stress.

Null Hypothesis 2: There is no significant relationship between the predictor (Public housing) and rent stress.

Alternative Hypothesis 2: There is a significant relationship between the predictor (Public Housing) and rent stress.

[2.2] Set-Up Evaluation

[2.2a] Feature Significance

To identify the most effective algorithm, various statistical models will be trained. If models such as Lasso, Ridge, or multiple linear regression yield the best results, p-values will be computed to observe their statistical significance.

For models such as Random Forest or XGBoosting, feature importance will be displayed to identify the most significant features. The project will go through iterations by applying feature engineering or selection techniques to assess the continued relevance of the variables. This process will provide valuable insights into the predictive power and importance of these variables.

[2.2b] Mean Squared Error

The best performer models will be based on the Mean Squared error (MSE) in the testing set, as it is well-suited for this regression project. The testing set will be used to assess the generalisation performance of a model and determine its effectiveness.

[2.2c] Confidence intervals

Confidence intervals at a 95% level will be calculated for all features, with particular emphasis on the variables used to assess the hypotheses, such as community housing and public housing. These confidence intervals will provide a range within which the coefficient could potentially lie. This analysis allows for a more comprehensive understanding of the uncertainty associated with the estimated coefficients.

[2.2d] P-testing / F-testing

In addition to calculating confidence intervals, a p-test will also be conducted to assess the hypothesis regarding the presence of a significant relationship between the predictor and rent stress. The p-value obtained will help determine whether to reject or accept the hypothesis. The significance level will be set in the approach phase given two-tailed hypothesis testing.

Section 3: Approach

[3.1] CRISP-DM

The project follows the "Cross-Industry Standard Process for Data Mining," which is a framework used for data mining in various fields. This process allows for regular evaluation and adjustment of project milestones as new information is discovered, insights are enhanced, and objectives are achieved.

[3.2] Cross-Validation

During the machine learning preparation phase, cross-validation remains crucial in practice to ensure reliable model performance. The data is divided into batches to facilitate training, testing, and validation of the model's output. Initially, the data is split into a training set **(80%)** and a testing set **(20%).**

The training dataset is then further divided into **multiple folds** using cross-validation. This allows the model to be trained and tested on different subsets of the data, helping to mitigate the risk of overfitting. The testing set is reserved for evaluating the model's performance.

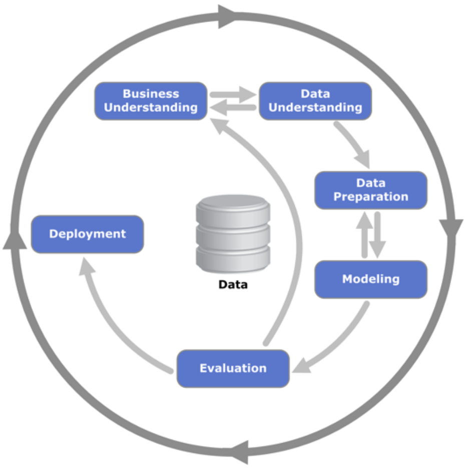


Figure 2 The CRISP-DM workflow.

Before making predictions using the model, it is important to standardise all the features. This is accomplished by extracting the mean of each feature and dividing it by the standard deviation. This standardisation process ensures that all features are on a similar scale, which can enhance the model's performance and interpretability.

[3.3] Model Assessment

To compare different machine learning models, the project will rely on the **Mean Squared Error** (MSE) values produced by each model. MSE is a regression analysis metric that quantifies the average squared difference between predicted and actual values. To calculate MSE, the predicted values are compared to the actual values in the testing set. Lower MSE values signify superior performance because they suggest that the model's predictions closely align with the actual values.

[3.4] Paired T-Test And Confidence Intervals

This project will use the **paired t-test** to evaluate whether there is a statistically significant difference in the mean of the testing MSE between the machine learning models. Performing a paired t-test involves comparing the means of two related groups or conditions where the same subjects are measured or tested under different conditions or at different time points. This statistical test is used to determine whether there is a significant difference between the means of paired observations.

The **null hypothesis** (H0) assumes that there is no difference in the test MSE mean between the models, while the **alternative hypothesis** (H1) states that there is a significant difference in the test MSE mean between the models. The testing MSE values for both models will be compared, and the p-value of the t-test will be computed. If the p-value will be below 0.05 the null hypothesis will be rejected.

Coefficient testing will be two-tailed. This means that the p-values to reject the null hypothesis will have to be below 0.025 (confidence level of 95%).

[3.5] Models Used For Regression Analysis

The approach will utilise linear models and regularised linear models for regression analysis. Specifically, it will focus on the implementation and evaluation of linear regression, Lasso (L1 regularised linear regression), Ridge (L2 regularised linear regression), K-Nearest Neighbors (KNN) regression, Random Forest regression and XGBoosting. The key steps will be the following:

Linear Regression, Lasso and Ridge:

Firstly, the team will fit a linear regression model using the `LinearRegression()` function and will evaluate the model's performance by computing mean squared error (MSE) on training and testing data.

The next step involves fitting a Lasso model and performing a grid search for hyperparameter tuning after defining a grid of hyperparameters, specifically the regularisation parameter alpha. The model's performance will be evaluated by computing MSE for different alpha values and plotting the results to compare training and testing MSE for Lasso regression. The final step is to fit a Ridge model and conduct a grid search to tune hyperparameters.

Description of hyperparameters that will be used:

In the Lasso regression, the regularisation parameter alpha controls the amount of shrinkage applied to the coefficients in Lasso regression. It determines the balance between the model's simplicity (smaller coefficient values) and its prediction accuracy. Grid search will be performed to explore different values of alpha and find the optimal value that minimises the mean squared error (MSE).

Similar to Lasso regression, the regularisation parameter alpha in Ridge regression controls the amount of shrinkage applied to the coefficients. It balances the model's complexity and prediction accuracy. Grid search will be performed to find the best value of alpha that minimises the MSE. The grid search technique is employed to systematically explore different values of the regularisation parameter within a defined range. By evaluating the model's performance using different alpha values, one can determine the optimal hyperparameter setting that yields the best predictive performance.

Model Comparison with a Paired T-Test

In order to compare the performance of the Lasso and Linear Regression models, the mean MSE will be calculated using cross-validation for both models. A paired t-test will be conducted to determine whether the mean difference of the test MSE between the two models is statistically significant.

Feature Selection with Lasso and Bootstrap resampling

A Lasso model with a specific alpha value will be instantiated to select the most relevant features. The Lasso model will then be fitted to the training data, the best predictors will be identified based on the non-zero coefficients, and the highest absolute coefficients will be determined.

Bootstrapping resampling will be performed by repeatedly fitting the Lasso model to bootstrap training data samples to determine whether community housing providers are essential features. The mean standard error, t-statistic, p-value, and confidence interval will be calculated for each coefficient.

K-Nearest Neighbours Regression

In order to check if the metrics can improve with a more flexible model K-Nearest Neighbours regression model (KNN) will be performed. The goal will be to understand if KNN could give insights on how flexible the model should be. Therefore, a grid of hyperparameters will be defined, including the 'n_neighbors' (the number of nearest neighbours to consider) and the 'metric' (the distance metric used to measure the similarity between instances). Subsequently, the KNN regression model will be fitted using the `KNeighborsRegressor()` function and perform a grid search for hyperparameter tuning. Finally, the model's performance will be evaluated by computing MSE for different hyperparameter values.

Random Forest Regression

For Random Forest Regression, a grid of hyperparameters will be defined, including the number of trees (n_estimators), maximum depth (max_depth), maximum features (max_features), and minimum samples per leaf (min_samples_leaf). a Random Forest regression model function will be fitted and the model's performance will be evaluated by computing MSE for different hyperparameter values. In order to compare the RF model with Lasso, the MSE for both the Random Forest and Lasso models will be calculated, and a statistical hypothesis test will be performed to determine if there is a significant difference in the test.

The hyperparameters that will be used are the following:

1. ***n_estimators***: This hyperparameter controls the number of decision trees in the random forest.
2. ***max_depth***: It determines the maximum depth allowed for each decision tree in the random forest.
3. ***max_features***: This hyperparameter specifies the number of features to consider when looking for the best split at each tree node.
4. ***min_samples_leaf***: this hyperparameter determines the minimum number of samples required to be at a leaf node of a decision tree.

These hyperparameters will be explored using grid search, where all possible combinations are evaluated to find the best set of hyperparameters that yield the optimal performance. The result will provide the best parameters found for the 'RandomForestRegressor' model, along with the training and testing mean squared errors (MSE) as evaluation metrics. The best parameters indicate the chosen values for each hyperparameter.

Random Forest Regressor and XGBRegressor

The random forest model is trained using the provided hyperparameters. Feature importance and corresponding feature names will be plotted in a bar chart, and the same process will be repeated for the top 15 features.

The model_cv function will be used to perform cross-validation with grid search, evaluating different combinations of hyperparameters. The results of the XGBRegressor model with different hyperparameter settings will be printed.

Model Comparison: Random Forest vs. Lasso

After the feature importance analysis for both models, a comparison will be made to determine if XGBRegressor performs better than Random Forest. The models are fit and tested using the fit_models function with the provided hyperparameters. The mean of the test mean squared error (MSE) for each model will be calculated and a paired t-test will be conducted to determine if there is a statistically significant difference in the test MSE means between the two models.

Feature Selection, Feature Engineering and Further Analysis

It will be necessary to select among the features to eliminate those that show a correlation when the models show signs of overfitting. Feature sets with high correlation can provide redundant information, so some can be eliminated.

In addition, feature engineering may be required to create some new variables to estimate the available spending capacity for both individuals and households after considering rent or mortgage payments and the cost of living. These new features can provide additional insights into the financial situation of individuals and households in the dataset.

Next, feature importance will be generated from the XGBoost and lasso models after performing hyperparameter tuning and model training. The coefficients' confidence intervals and p-values will be estimated by bootstrapping resampling, and the features with p-values less than 0.025 will be considered.

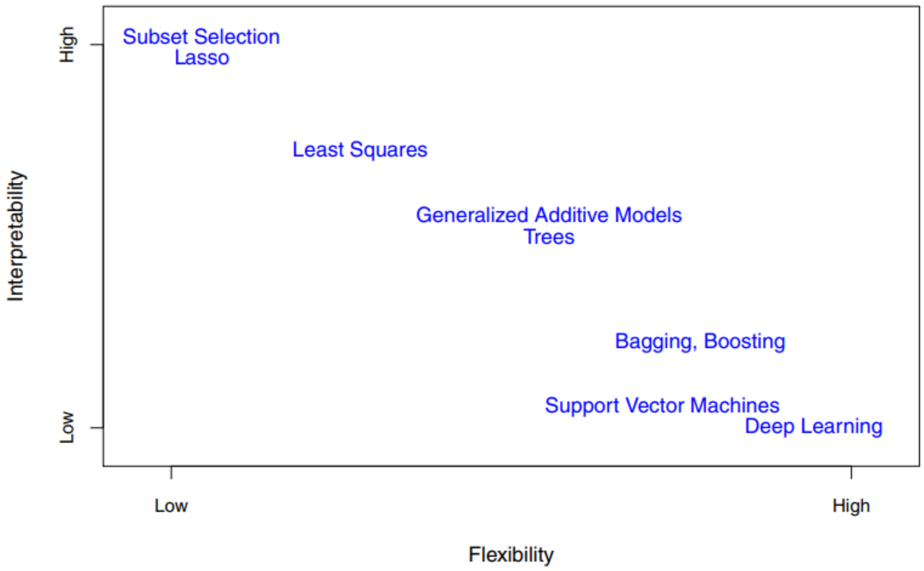


Figure 3 Extract from the book "An Introduction to Statistical Learning with Applications in R - Second Edition."

Section 4: Data

Following construction and training a range of regression models, including:

- Multiple Linear
- Lasso
- K-Nearest Neighbors
- Random Forest
- XG Boosting Regressor (*with decision trees as a base*);

.. preliminary observations revealed the first two models experienced no overfitting problems, while the latter three, unfortunately, had. To address overfitting within these models, regularisation via hyperparameter tuning was first trialled.

Despite applying multiple iterations of hyperparameter variations, the difference gap between the performance of each model, in regards to training and testing outcomes, remained broad and dissimilar. Investigations that imposed stricter regularisation with reduced values for specific hyperparameters such as “max_depth” and “max_features” (*which represent a subsample for the XGBoosting iteration; for a complete list of hyperparameters explored and their corresponding model, see appendix*), yielded a marginally improved outcome for the corresponding training Mean Squared Error (MSE); though these refinements still remained inadequate and non-conclusive.

As adjustments to hyperparameter inclusions and their corresponding ranges yielded unsatisfactory outcomes, feature engineering and determining feature specification were next explored as a means to elucidate any variable significance. The approach was further prioritised, as engineering variables would help clarify any explanatory influence of some features over others. Specifying feature inclusions to model, on the other hand, would serve to help reduce overfitting.

While the project objective was to explore insights about the more significant features, a strong MSE testing outcome was also prioritised, as it indicates whether a model has understood the data. This is valuable - when pulled off successfully, the finalised model could be applied to generalise new datasets about housing or geographical zones. A model that generalises typically helps explain which feature variance relates to the target variable.

The feature specification process involved dropping highly correlated features (coefficients greater than 0.85), in addition to geographical features. These are summarised as followed:

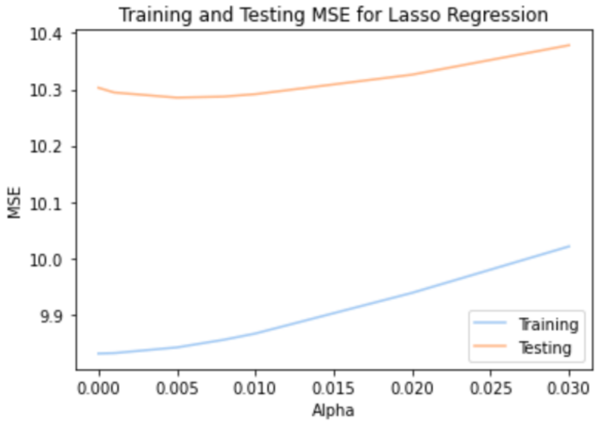
Baseline model

Construction of a baseline model was carried out as a means to predict the average target variable within each observation. Our baseline values revealed a training MSE of 94.5 and a testing MSE of 92.5, indicating an overall poor model performance and a high degree of variance with each output. The RMSE values for the training and testing data sets were 9.72 and 9.62, respectively. With a 95% confidence interval, the baseline's prediction interval of 19.24 indicated a significant error. To further contextualise this degree of error, the interquartile range (IQR) for the target variable was calculated as 12.85.

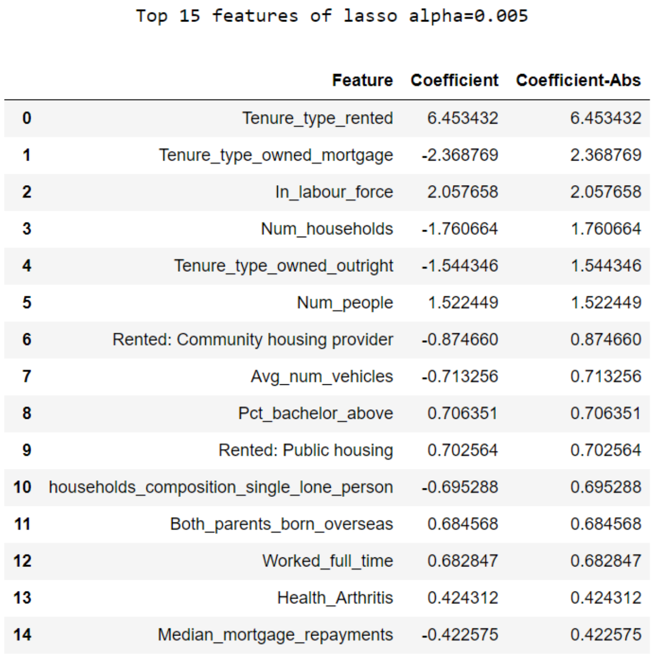
Multiple Linear Regression & Lasso

Modelling via linear regression revealed a training MSE of 9.8658 and a testing MSE of 10.3781. The results suggest that the model may slightly overfit the dataset, as the training MSE was lower than the testing MSE.

A Lasso regression was subsequently pursued in effort to lower possible variance. Despite this, the model only managed to lower the MSE on average, by approximately 0.07 - a variation not considered impactful enough in context of project's goals. The most suited parameter to adjust was determined as alpha, where optimal ranges resided between 0.005 and 0.008, as depicted by the plot 1.



An advantage of applying multiple linear & Lasso regression models is the added functionality to calculate supplementary metrics, such as coefficients, p-values, and confidence intervals for each feature.
Based on the ranges of the coefficients in the dataset, the Lasso regression model (*alpha: 0.005*) was determined as the best predictor and bore strongest associated features with the target variable. As the p-values reflect the statistical significance of coefficients, while the confidence intervals estimate the plausible range of values for the coefficients (calculated with 95% confidence), these metrics were also examined. Their integration into the study also granted insights into the relationship between the features and target variable, the model's overall performance, and identification into potential sources of bias.



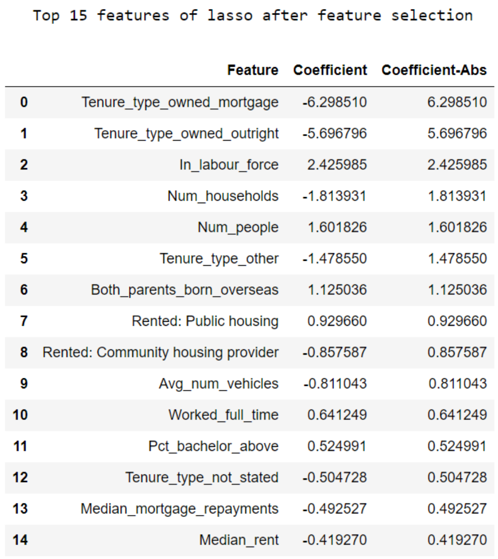
```
Top 15 features of lasso alpha=0.005
```

	Feature	Coefficient	Coefficient-Abs
0	Tenure_type_rented	6.453432	6.453432
1	Tenure_type_owned_mortgage	-2.368769	2.368769
2	In_labour_force	2.057658	2.057658
3	Num_households	-1.760664	1.760664
4	Tenure_type_owned_outright	-1.544346	1.544346
5	Num_people	1.522449	1.522449
6	Rented: Community housing provider	-0.874660	0.874660
7	Avg_num_vehicles	-0.713256	0.713256
8	Pct_bachelor_above	0.706351	0.706351
9	Rented: Public housing	0.702564	0.702564
10	households_composition_single_lone_person	-0.695288	0.695288
11	Both_parents_born_overseas	0.684568	0.684568
12	Worked_full_time	0.682847	0.682847
13	Health_Arthritis	0.424312	0.424312
14	Median_mortgage_repayments	-0.422575	0.422575

The leading 15 features of the dataset that contribute to rent stress in NSW have ultimately been calculated and summarised through a Lasso regression (*alpha: 0.005*), to address the primary research goal in determining which variables or characteristics contribute to NSW's rent stress. The leading predictors of rent stress have been determined through the highest absolute association of the target variable with each feature within the dataset.

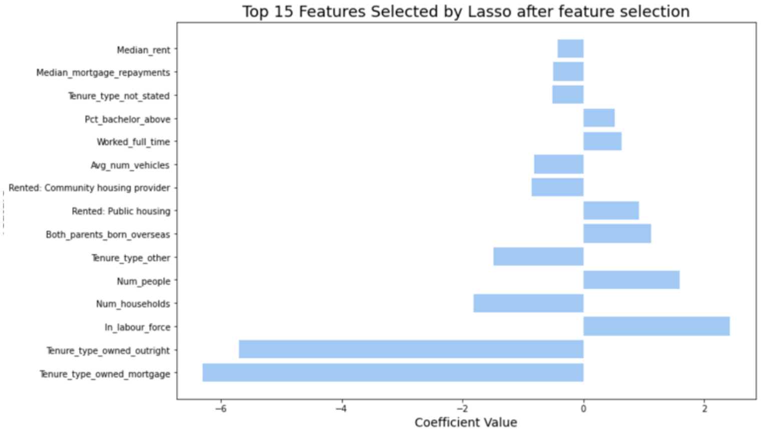
The resulting data frame revealed that the most prominent feature among all the columns in the dataset is "Tenure_type_rented" - representing the percentage of households renting during the 2021 census. As rent stress has a known prevalence in regions with a higher percentage of rented households, this feature was not deemed informative and provided no value to relevant stakeholders. For this reason, "Tenure_type_rented" was excluded from further analysis.

Following feature engineering (*creating new features titled "spending freedom", or "spending capacity" following rental payment and associated costs of living*), and feature specification/discontinuation through the removal of highly correlated features (coefficient values above 0.85), the leading 15 features remained unchanged. However, the change was considered most significant the moment "Tenure_type_rented" was discontinued, as the Lasso model's test MSE performance considerably dropped. While the outcome is deemed unfavourable, these results paved a new perspective, as it emphasised the value of "Tenure_type_owned_mortgage" and "Tenure_type_owned_outright" as significant predictors.



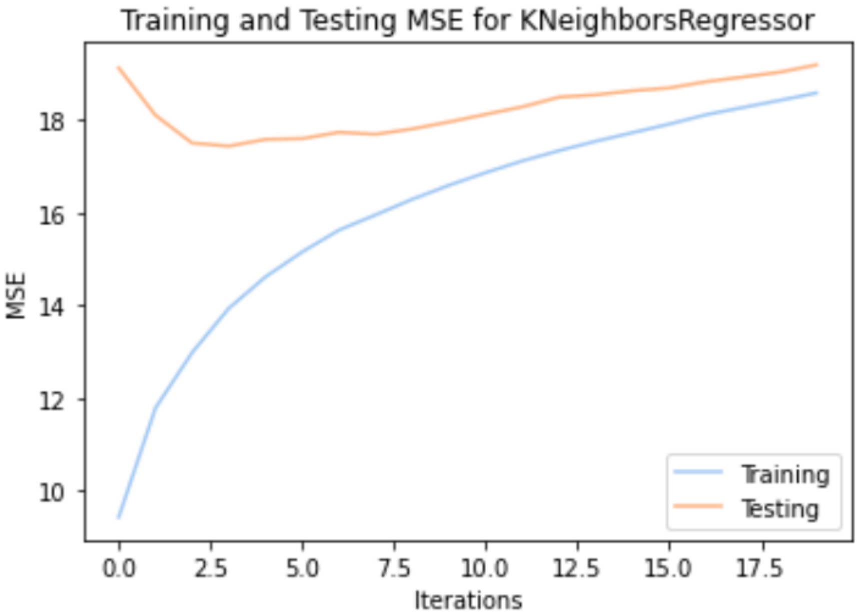
```
Top 15 features of lasso after feature selection
```

	Feature	Coefficient	Coefficient-Abs
0	Tenure_type_owned_mortgage	-6.298510	6.298510
1	Tenure_type_owned_outright	-5.696796	5.696796
2	In_labour_force	2.425985	2.425985
3	Num_households	-1.813931	1.813931
4	Num_people	1.601826	1.601826
5	Tenure_type_other	-1.478550	1.478550
6	Both_parents_born_overseas	1.125036	1.125036
7	Rented: Public housing	0.929660	0.929660
8	Rented: Community housing provider	-0.857587	0.857587
9	Avg_num_vehicles	-0.811043	0.811043
10	Worked_full_time	0.641249	0.641249
11	Pct_bachelor_above	0.524991	0.524991
12	Tenure_type_not_stated	-0.504728	0.504728
13	Median_mortgage_repayments	-0.492527	0.492527
14	Median_rent	-0.419270	0.419270



K-Nearest Neighbours

The performance of this statistical learning model was highly unsatisfactory, as indicated by the Training Mean Squared Error (MSE) of 13.937 and the Testing MSE of 17.436. It significantly underperformed even in comparison to an inflexible model such as multiple linear regression. Despite attempts to optimise the model's hyperparameters, it consistently failed to achieve a Testing MSE below 17.

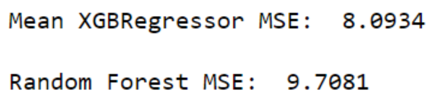


Random Forest Regression

Despite showing signs of overfitting, the model's performance was marginally better than that of the Lasso model from a statistical standpoint. However, the practical implications for the project were similar in terms of the impact and significance of the difference.

GXBoostingRegressor

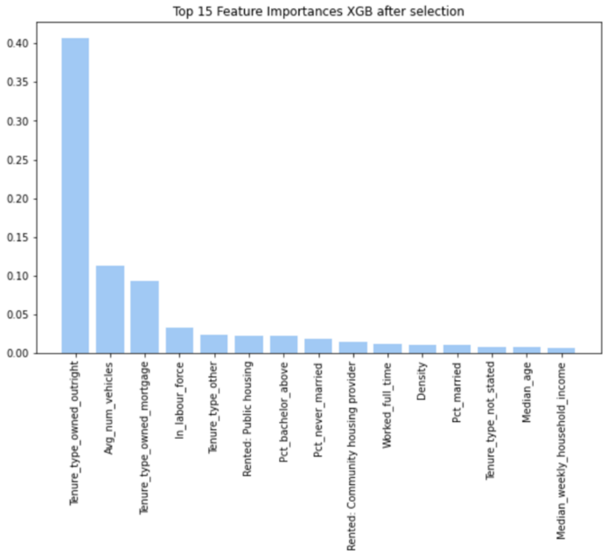
While the Lasso regression’s interpretation was considered straight forward, it’s performance wasn’t strong. It was next determined that applying the GXBoosting Regressor library should be explored, due to its flexibility and boosting functionality through decision trees. With this decision, a much lower MSE value across training and testing sets (than either KNN or Random Forest regression, despite tuning hyperparameters for both) was attained. The final testing MSE across 10 iterations, was 8.09 (see figure 2).

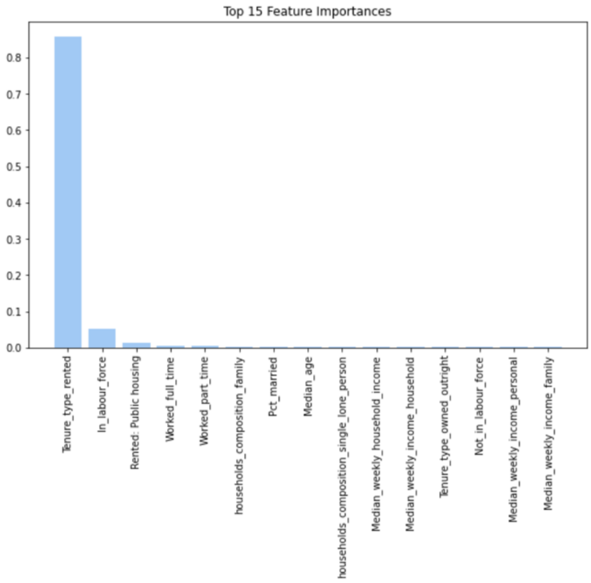


```
Mean XGBRegressor MSE:  8.0934

Random Forest MSE:  9.7081
```

Using the “feature importance” functionality, integrated within XGBoost library, the most important features could be plotted (Ye et al., 2019, figure 3). Once more, as "Tenure_type_rented" contributed no insights towards determining rent stress and providing value to stakeholders, it was discontinued from the line of features, resulting in figure 4.





“Is there an association between government interventions and a reduction

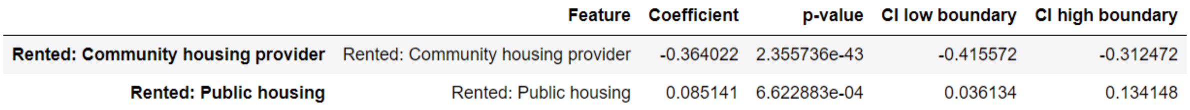
of rent stress in New South Wales?”

The second research inquiry that was addressed in this project aimed to determine if there was a relationship between the government interventions and the reduction of rent stress in New South Wales, based on significant evidence.

To investigate this question, the predictors "Rented: Community housing provider" and "Rented: Public Housing" were examined separately to determine if there was a statistically significant relationship between these features and a lower percentage of Household rent stress. A least squares regression method was utilised to calculate the strength and significance of the relationship, followed by Lasso. Confidence intervals and p-values were then computed.

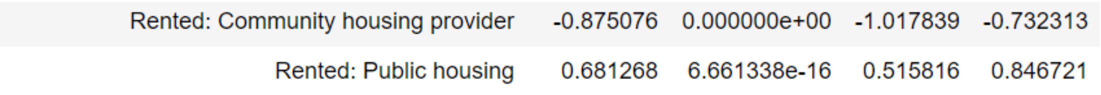
The government has two primary means of offering affordable housing options. The first approach is via public housing, which involves full government management. Alternatively, the government can fund and regulate community housing providers. In both scenarios, the government plays a crucial role in facilitating access to affordable housing.

According to the multiple linear regression analysis, both "Rented: Community housing provider" and "Rented: Public Housing" were statistically significant with a 95% confidence level. However, the analysis was expected to reveal a negative relationship for both, but a positive relationship was observed for Public housing.



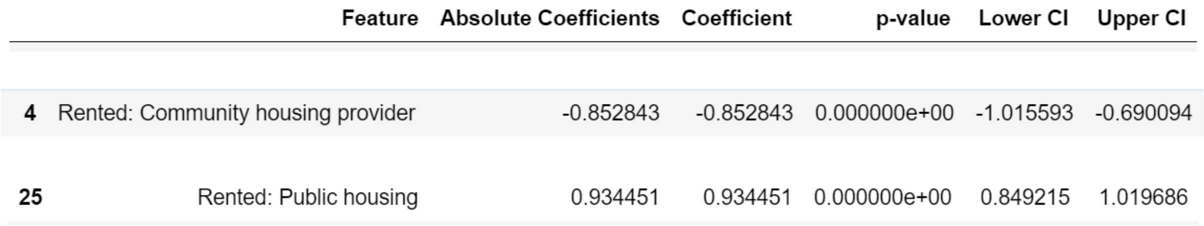
	Feature	Coefficient	p-value	CI low boundary	CI high boundary
Rented: Community housing provider	Rented: Community housing provider	-0.364022	2.355736e-43	-0.415572	-0.312472
Rented: Public housing	Rented: Public housing	0.085141	6.622883e-04	0.036134	0.134148

The Lasso model produced consistent results in terms of direction and significance for both "Rented: Community housing provider" and "Rented: Public Housing" features. However, in this model, both features exhibited a steeper slope, indicating their higher significance in predicting the outcome variable. Bootstrapping with replacement was utilized to calculate the mean and variance of the coefficients. These statistics are essential for computing the t statistic, p-values, and confidence intervals.



Rented: Community housing provider	-0.875076	0.000000e+00	-1.017839	-0.732313
Rented: Public housing	0.681268	6.661338e-16	0.515816	0.846721

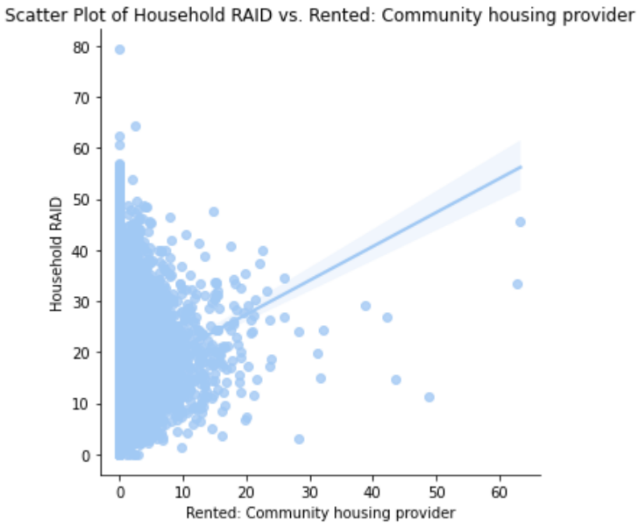
After feature engineering, feature selection without ternure type rented a Lasso was fitted and also computed p-values and confidence intervals through bootstrapping.



	Feature	Absolute Coefficients	Coefficient	p-value	Lower CI	Upper CI
4	Rented: Community housing provider	-0.852843	-0.852843	0.000000e+00	-1.015593	-0.690094
25	Rented: Public housing	0.934451	0.934451	0.000000e+00	0.849215	1.019686

Both hypotheses can be confidently rejected with 95% and even 99% confidence, as the p-values are essentially zero. It was initially expected to observe inverse relationships in both cases. However, in the case of public housing, it was found that an increase in the presence of public housing in a geographical area is associated with a higher percentage of households experiencing rent stress.

During the Exploratory Data Analysis, the simple linear relationship showed a positive association as depicted in the next graph. The insight drawn from the scatter plot was that community housing had a higher presence in areas where rent stress was more pronounced.



However, after accounting for confounding variables, the relationship changed direction. It became evident that community housing providing affordable housing options was associated with lower levels of rent stress. When controlling for lurking values of coefficients changed from .066 positive in case of simple linear regression to -0.85 for lasso alpha 0.005 after feature selection.

Section 5: Conclusion

The objective of this project was to address two main questions related to rent stress in New South Wales:

"What variables or characteristics contribute to rent stress in New South Wales?"

"Is there an association between government interventions and a reduction of rent stress in New South Wales?"

Several regression models were fitted to answer these questions, including:

Multiple Linear Regression,
Lasso,
KNN,
Random Forest, and
XGBoosting. Among these models, XGBoosting emerged as the top performer. Using the feature importance method from scikit-learn, the project identified and plotted the top 15 variables contributing to rent stress.

This analysis was performed with three iterations:

1. The first iteration included all the features of the model,
2. The second iteration was conducted after feature engineering and selection,
3. The third iteration involved eliminating the tenure-type feature.

In order to answer the first research question, the features that appeared in at least two of the three iterations in XGBoosting as significant predictors of rent stress in New South Wales are:

In_labour_force
Tenure_type_rented
Rented: Public housing
Tenure_type_owned_outright
Worked_full_time
Rented: Community housing provider
Median_age
Density
Median_weekly_household_income
Pct_married
Worked_part_time

These are essential parameters for the models that best learnt about the problem. They displayed the lowest generalisation error (Train MSE) over the rest.
The second question was addressed by assessing the statistical significance of the variables using the hypothesis testing approach. This involved examining the p-values and confidence intervals to determine their significance level.

However, XGBoosting and random forest algorithms only provide feature importance without any statistical computations to determine whether to accept or reject the null hypothesis. In order to determine the statistical significance of the coefficients related to government efforts, multiple linear regression and Lasso were used. In all iterations of both methods, the coefficients consistently displayed p-values for those features well below 0.025, leading to the rejection of both null hypotheses. This indicates **a significant association between the features and rent stress, the predicted variable**. However, it should be noted that only in the case of "Rented: Community housing" is the relationship inverse.

Based on the final Lasso model selection, it can be interpreted as the association between "Rented: Community housing" and rent stress as follows: for every 1% increase in community housing as a percentage of total households within a zone, there was an average decrease of 0.85% in rent stress. The actual population coefficient lies between 1.01% and 0.69% with 95% confidence.
The government's efforts in providing public housing were found to have a statistically significant association with rent stress, as indicated by a confidence interval of 0.85 to 1.01 in a positive direction. This suggests the alternative hypothesis can be accepted, indicating a clear difference from zero.
It is essential to acknowledge that causation cannot be inferred in an observational study such as this. However, the study provides valuable insights into the potential inefficiency of public housing in reducing rent stress. Observational studies can serve as a foundation for designing future experiments that aim to determine causation, therefore, the effectiveness of public housing. In such experimental designs, controlling for additional variables and studying at an individual level rather than geographical areas would be crucial. Despite the limitations of the study, it is evident that there is a strong association between Public and Community housing and rent stress.

References

Literature Review

[1.1] **Ye, T., Johnson, R., Fu, S., Copeny, J., Donnelly, B., Freeman, A., Lima, M., Walsh, J., & Ghani, R.** (2019). *Using Machine Learning to Help Vulnerable Tenants in New York City.*

[1.2] **Muthukrishnan, R., & Rohini, R.** (2016, October). *LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 18-20). IEEE.

[1.3] **Keith Jacobs, Rowland Atkinson, Val Colic Peisker, Mike Berry and Tony Dalton. (2010, September).** *What future for public housing? A critical analysis.* **AHURI. Report. 151.**

[1.4] **Tan, J.** (2020). Using Machine Learning to Identify Populations at High Risk for Eviction as an Indicator of Homelessness. (Master's thesis). Massachusetts Institute of Technology.

[1.5] **Muir, K., Moran, M., Michaux, F., Findlay, S., Meltzer, A., Mason, C., ... & Heaney, R. A. (**2017). The opportunities, risks and possibilities of social impact investment for housing and homelessness.

Appendix:

Dataset Features:

0. **Num_people**: The number of people in the area.
1. **Pct_female**: The percentage of females in the area.
2. **Median_age**: The median age of the population in the area.
3. **Num_households**: The number of households in the area.
4. **Avg_people_per_households**: The average number of people per household.
5. **Median_weekly_household_income**: The median weekly household income.
6. **Median_monthly_mortgage_repayments**: The median monthly mortgage repayments.
7. **Median_weekly_rent**: The median weekly rent.
8. **Avg_num_vehicles**: The average number of vehicles per household.
9. **Pct_indigenous**: The percentage of indigenous population.
10. **Pct_non_indigenous**: The percentage of non-indigenous population.
11. **Pct_Non_stated_people**: The percentage of people who did not state their ethnicity.
12. **Pct_married**: The percentage of married people.
13. **Pct_separated**: The percentage of separated people.
14. **Pct_divorced**: The percentage of divorced people.
15. **Pct_widowed**: The percentage of widowed people.
16. **Pct_never_married**: The percentage of people who have never been married.
17. **Pct_bachelor_above**: The percentage of people with a bachelor's degree or above.
18. **Pct_advanceddiploma_diploma**: The percentage of people with an advanced diploma or diploma.
19. **Pct_CertIV**: The percentage of people with a Certificate IV qualification.
20. **Pct_certIII**: The percentage of people with a Certificate III qualification.
21. **Year_12**: The percentage of people who completed Year 12 education.
22. **Year_11**: The percentage of people who completed Year 11 education.
23. **Year_10**: The percentage of people who completed Year 10 education.
24. **Pct_CertII**: The percentage of people with a Certificate II qualification.
25. **Pct_CertI**: The percentage of people with a Certificate I qualification.
26. **Pct_year_9_below**: The percentage of people with education below Year 9 level.
27. **Pct_no_educational_attainment**: The percentage of people with no educational attainment.
28. **Both_parents_born_overseas**: The percentage of people with both parents born overseas.
29. **Father_only_born_overseas**: The percentage of people with only their father born overseas.
30. **Mother_only_born_overseas**: The percentage of people with only their mother born overseas.
31. **Both_parents_born_ustralia**: The percentage of people with both parents born in Australia.
32. **In_labour_force**: The percentage of people in the labor force.
33. **Not_in_labour_force**: The percentage of people not in the labor force.
34. **Worked_full_time**: The percentage of people who worked full time.
35. **Worked_part_time**: The percentage of people who worked part time.
36. **Unemployed**: The percentage of unemployed people.
37. **Median_weekly_income_personal**: The median weekly personal income.
38. **Median_weekly_income_family**: The median weekly family income.
39. **Median_weekly_income_household**: The median weekly household income.
40. **households_composition_family**: The percentage of households composed of families.
41. **households_composition_single_lone_person**: The percentage of households composed of single/lone persons.
42. **households_composition_group**: The percentage of households composed of groups.
43. **Tenure_type_owned_outright**: The percentage of households with outright ownership.
44. **Tenure_type_owned_mortgage**: The percentage of households with ownership through a mortgage.
45. **Tenure_type_rented**: The percentage of households that are rented.
46. **Tenure_type_other**: The percentage of households with other tenure types.
47. **Tenure_type_not_stated**: The percentage of households with unstated tenure type.
48. **Median_rent**: The median rent.
49. **Median_mortgage_repayments**: The median monthly mortgage repayments.
50. **Household** RAID: The household Relative Advantage and Disadvantage (RAID) index.

51. **Avg_people_per_bedroom_aboriginal**: The average number of people per bedroom for Aboriginal population.
52. **Median_weekly_household_income_aboriginal**: The median weekly household income for Aboriginal population.
53. **Health_no_condition**: The percentage of people with no reported health conditions.
54. **Health_one_condition**: The percentage of people with one reported health condition.
55. **Health_two_conditions**: The percentage of people with two reported health conditions.
56. **Health_three_more_conditions**: The percentage of people with three or more reported health conditions.
57. **Health_Arthritis**: The percentage of people with arthritis.
58. **Health_Asthma**: The percentage of people with asthma.
59. **Health_Cancer**: The percentage of people with cancer.
60. **Health_Dementia**: The percentage of people with dementia.
61. **Health_Diabetes**: The percentage of people with diabetes.
62. **Health_Heart_disease**: The percentage of people with heart disease.
63. **Health_Kidney_disease**: The percentage of people with kidney disease.
64. **Health_Lung**: The percentage of people with lung disease.
65. **Health_Mental_condition**: The percentage of people with mental health conditions.
66. **Health_Stroke**: The percentage of people who have had a stroke.
67. **Health_other_conditions**: The percentage of people with other reported health conditions.
68. **Health_no_longterm_conditions**: The percentage of people with no reported long-term health conditions.
69. **Car_driver**: The percentage of people who are car drivers.
70. **Car_passenger**: The percentage of people who are car passengers.
71. **SA1 Code**: The statistical area 1 code.
72. **Area sqkm**: The area in square kilometers.
73. **Region**: The region of the area.
74. **State**: The state of the area.
75. **SA4_NAME_2016**: The statistical area 4 name in 2016.
76. **Density**: The population density.
77. **Rented**: Community housing provider: The percentage of households rented from a community housing provider.
78. **Rented**: Public housing: The percentage of households rented from public housing.
79. **Households rented CH or PH**: The number of households rented from community housing or public housing.

Model Outcomes:

Baseline:

Baseline training MSE: 94.5
Baseline testing MSE: 92.5
Baseline training RMSE: 9.72111104761179
Baseline testing RMSE: 9.617692030835672

Linear regression:

Best Parameters: N/A
Training MSE: 9.8658
Testing MSE: 10.3781

Lasso L1 Regularisation:

Best Parameters: {'alpha': 0.005}
Average Lasso MSE: 10.0623
Average Linear Regression MSE: 10.076400000000003
t-test p-value: 0.0003413713779647204

Ridge L2 Regularisation:

Best Parameters: {'alpha': 20}
Training MSE: 9.8325
Testing MSE: 10.2934

KNN:

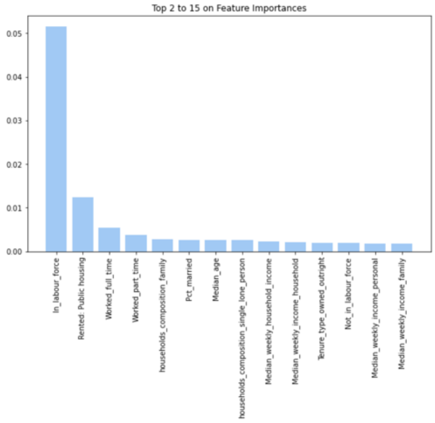
Best Parameters: {'metric': 'manhattan', 'n_neighbors': 9}

K:	9.000000
Training:	13.937817
Testing:	17.436036

Random Forest Regression:

Best Parameters:
{'max_depth': 20, 'max_features': 'auto', 'min_samples_leaf': 10,
'min_samples_split': 30, 'n_estimators': 500}

Training MSE: 5.541
Testing MSE: 10.0651
t-test p-value: 0.002148531643802524



model	Lasso	RandomForestRegressor
random_state		
0	10.359	10.352
1	9.699	9.615
2	10.066	9.685
3	9.449	9.242
4	9.933	9.721

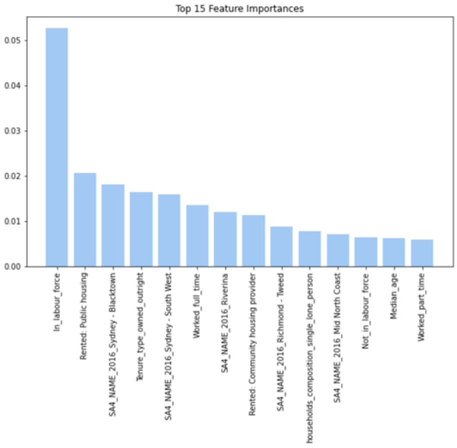
Feature	Coefficient	Coefficient-Abs	
45	Tenure_type_rented	6.453432	6.453432
44	Tenure_type_owned_mortgage	-2.368769	2.368769
32	In_labour_force	2.057658	2.057658
3	Num_households	-1.760664	1.760664
43	Tenure_type_owned_outright	-1.544346	1.544346
0	Num_people	1.522449	1.522449
72	Rented: Community housing provider	-0.874660	0.874660
8	Avg_num_vehicles	-0.713256	0.713256
17	Pct_bachelor_above	0.706351	0.706351
73	Rented: Public housing	0.702564	0.702564
41	households_composition_single_lone_person	-0.695288	0.695288
28	Both_parents_born_overseas	0.684568	0.684568
34	Worked_full_time	0.682847	0.682847
56	Health_Arthritis	0.424312	0.424312
49	Median_mortgage_repayments	-0.422575	0.422575

Feature	Coefficient	p-value	CI low boundary	CI high boundary	
Rented: Community housing provider	Rented: Community housing provider	- 0.364022	2.355736e-43	-0.415572	- 0.312472
Rented: Public housing	Rented: Public housing	0.085141	6.622883e-04	0.036134	0.134148

XGBoost Regression:

Best Parameters:
{'gamma': 0.7, 'learning_rate': 0.1, 'max_depth': 4,
'n_estimators': 500, 'subsample': 0.6}

Training MSE: 3.6661
Testing MSE: 8.3308


Top 15 Feature Importances
In_labour_force
Rented: Public housing
SA4_NAME_2016_Sydney - Blacktown
Tenure_type_owned_outright
SA4_NAME_2016_Sydney - South West
Worked_full_time
SA4_NAME_2016_Riverina
Rented: Community housing provider
SA4_NAME_2016_Richmond - Tweed
households_composition_single_lone_person
SA4_NAME_2016_Mid North Coast
Not_in_labour_force
Median_age
Worked_part_time