**Assignment 1(C)**

**PART 1: Problem Description**

To derive a gentrification classification, Socio-Economic Indexes for Areas (SEIFA), set of indexes developed by the Australian Bureau of Statistics (ABS) is utilized to measure and rank the relative socio-economic advantage and disadvantage of different regions across Australia. According to Reades et al. (2019), it is vital to calculate the changes based on relative rather than absolute measurement because different suburbs might experience changes at different rates and absolute measures disregard the starting point of the suburb before the observed changes happened.

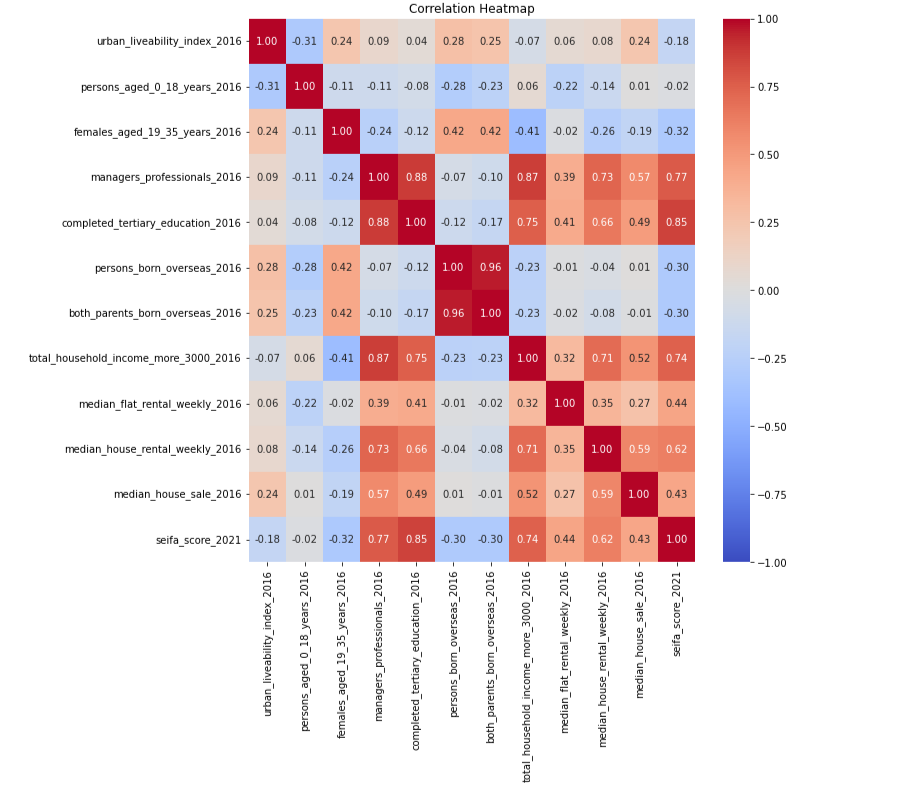


Figure : Correlation Matrix between 2016 input variable to 2021 SEIFA score.

Figure 1 shows the correlation coefficients between the **'seifa\_score\_2021**’ and other input features. The features that have notable positive correlations with the **'seifa\_score\_2021'** were **'completed\_tertiary\_education\_2016', 'managers\_professionals\_2016', 'total\_household\_income\_more\_3000\_2016', 'median\_house\_rental\_weekly\_2016', and 'median\_flat\_rental\_weekly\_2016**'.

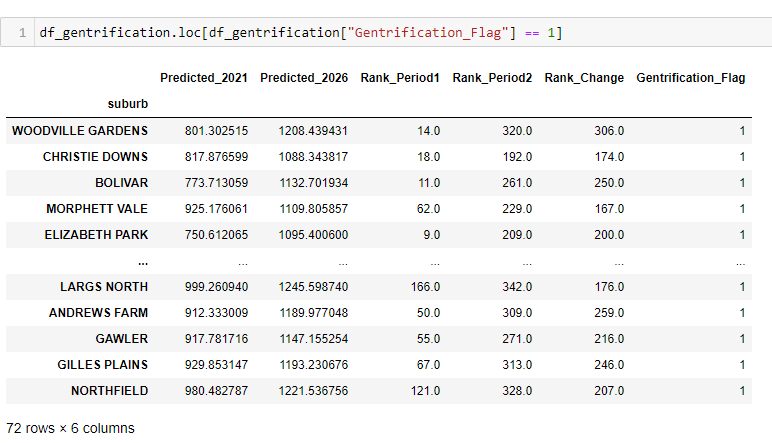


Figure : Predicted Suburbs within 2021 and 2026

The ranking process was adopted from Thackway et al. (2023) approach. Suburbs were arranged highest to lowest according to their differences in SEIFA score. If the difference above one standard deviation, the suburb was classified as 'gentrifying' while within the range of one standard deviation were classified as 'not gentrifying' (Thackway et al. 2023, p.5). One constraint of this approach was that suburbs with small changes, even though they already have a high SEIFA score, would not be captured (Thackway et al. 2023, p.5). This approach was mainly concerned to explore suburbs that would experience early gentrification rather than already gentrified (Thackway et al. 2023, p.5). The concept was to predict SEIFA 2021 score using input variables of 2016. With the model that had been tuned, unknown 2026 SEIFA score for Adelaide suburb areas was predicted using 2021 as the input variables. It is important to mention that this study aims to investigate potential changes in neighbourhoods from 2021 onwards, focusing on future transformations rather than comprehensively analysing the entire process of gentrification, which is very complex.

**Part 2: Data Pre-processing**

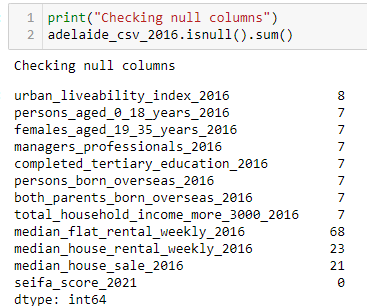
****

Figure : Checking null data.

Data pre-processing was performed to check if the dataset contains null data as per Figure 3.

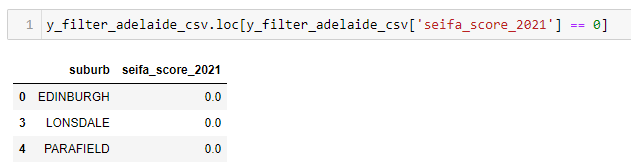


Figure : Filtered 2021 SEIFA score of 0.

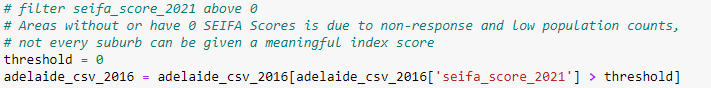
****

Figure : Removed SEIFA score of 0.

According to Figure 4 & 5, it is possible for suburbs to have null or 0 SEIFA Score index due to non-response and low population counts. Hence, suburbs that have 0 or Null had been removed.

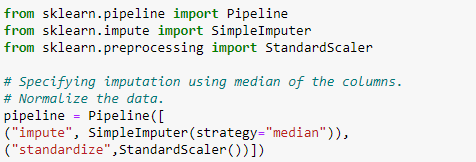
****

Figure : Pipeline in the pre-processing.

Figure 6 shows a pipeline had been initialised with imputer to impute any null data in the input columns while standard scaler was to normalise the dataset to have a similar range or distribution.

**Part 3: Model Selection**

The Scikit Learn library provide various machine learning model and randomized train/test indices to split data accordingly. The 2016 input dataset had been split into training and testing sets contained 275 and 118 points respectively. The model was trained using the 2016 input variables to predict the 2021 SEIFA index. A simple linear regression, Support Vector Regressor (SVR) and Kernel Ridge Regressor (KRR) are being utilized. For the SVR and KRR to perform at their best, each model would be optimised by utilising several hypermeters to maximise model performance.

**Part 4: Model Refinement**

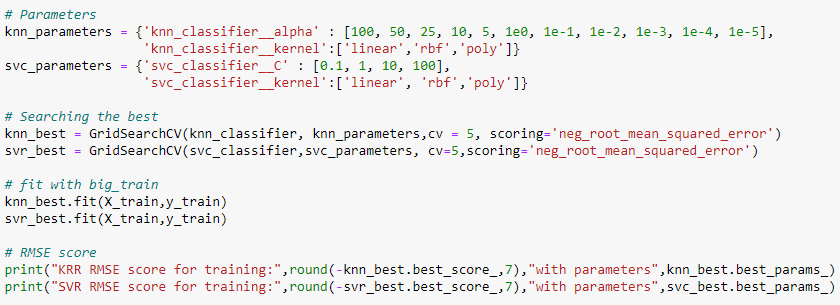


Figure : Model refinement process.

As per Figure 7, both models were using a similar number of hyperparameters for two reasons. First, to achieve simpler model selection by reducing the complexity of the searching the best parameters. Second, having fewer hyperparameters to tune can help prevent overfitting. Alas, the model can provide better generalization to new unseen data while avoiding overfitting. Each model method requires a different set of hyperparameters; hence they were tuned sequentially using cross-validation. The grid search was performed and fitted the value given the variability of the parameters. The model performance gets the average performance of each cross validation and the value with the highest average score was selected. The score was obtained from the best was measured using Root Mean Square Error (RMSE), The lower the RMSE value, indicating that model measurement was closer to the true values.

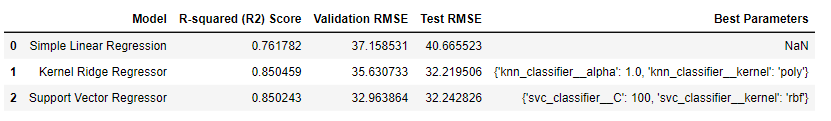
**Part 5: Performance Description**

Figure 9: Model performances.

Figure 8: Scatterplot chart between Predicted and Truth values.

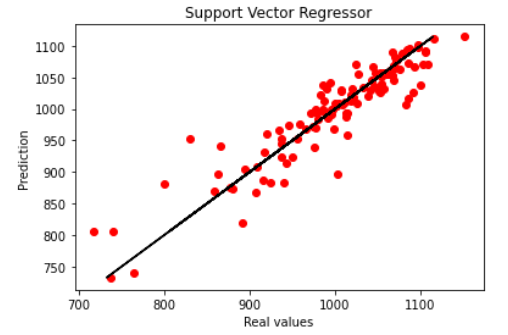
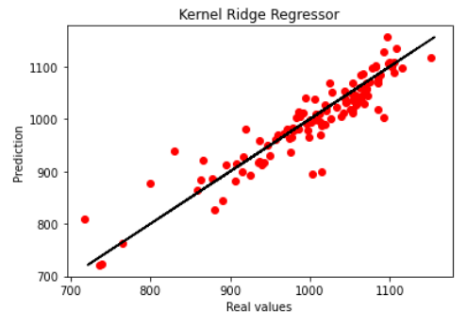
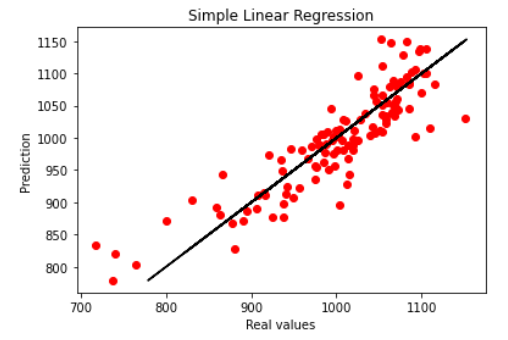


Figure 8 & 9 displays model performances that were evaluated based on three metrics: R-squared (R2) Score, validation RMSE, and testing RMSE after KRR and SVR models had been tuned. The highest R2 score was achieved by the KRR, followed by the SVR and the Simple Linear Regression. Both the SVR and KRR perform similarly well.

**Part 6: Results Interpretation**

KRR model was the most appropriate for selection because it achieved the highest R2 score, and outperformed the other models based on both validation and testing set, indicating that was a good fit to the data. Low RMSE in testing set shows that KRR provide better generalization to new unseen data while avoiding overfitting. For KRR, the choice of the **polynomial** kernel with **alpha** value 1.0 indicates the model's ability to capture nonlinear relationships in the data. For KRR, the hyperparameters that need to be optimized are 'alpha' and 'kernel'. The 'alpha' hyperparameter controls the regularization strength in the KRR model, helping to prevent overfitting. The 'kernel' hyperparameter was transforming the original input features into a higher-dimensional space, where the data may become linearly separable.

**References**

Badcock, B 2001, ‘Thirty years on: Gentrification and class changeover in Adelaide’s inner suburbs, 1966-96’, *Urban Studies (Edinburgh, Scotland)*, vol. 38, no. 9, pp. 1559–1572.

Baker, E 1997, ‘Gentrification and gender effects in North Adelaide’, Thesis (B.A.(Hons.)) --University of Adelaide, Dept. of Geography, 1998.

Kolko, J 2010, ‘The determinants of gentrification’*, SSRN Working paper Series*.

Loukaitou-Sideris, A, Gonzalez, S and Ong, P 2017. ‘Triangulating Neighborhood Knowledge to Understand Neighborhood Change: Methods to Study Gentrification’, *Journal of Planning Education and Research*, 39(2), pp.227–242, DOI:10.1177/0739456x17730890.

Palafox, L & Ortiz-Monasterio, P 2020, ‘Predicting Gentrification in Mexico City using Neural Networks’, in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–5.

Reades, J, De Souza, J & Hubbard, P 2019, ‘Understanding urban gentrification through machine learning’, *Urban Studies (Edinburgh, Scotland)*, vol. 56, no. 5, pp. 922–942.

Thackway, W, Ng, M, Lee, C-L & Pettit, C 2023, ‘Building a predictive machine learning model of gentrification in Sydney’, *Cities*, vol. 134, p. 104192–.

Zeng, J, Yue, Y, Gao, Q, Gu, Y & Ma, C 2022, ‘Identifying localized amenities for gentrification using a machine learning-based framework’, *Applied Geography (Sevenoaks)*, vol. 145, p. 102748–.