Assignment 1  
Regression Models

short line

Md Saifur Rahman

Student ID: 25528668  
2025 AUTUMN

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Business Understanding 2**](#_if9iwijz2pg3)

[a. Business Use Cases 2](#_bnjfuh5ewogw)

[**2. Data Understanding 3**](#_yg03y3w5hig9)

[**3. Data Preparation 4**](#_tj77rrmixzhi)

[**4. Modeling 5**](#_6vwh14nc44ga)

[**5. Evaluation 6**](#_mngiyx5tet6w)

[a. Results and Analysis 6](#_25v57zsj5m7g)

[b. Business Impact and Benefits 6](#_hkob76wu4d6q)

[c. Data Privacy and Ethical Concerns 6](#_uf1z6gbsejg6)

[**6. Conclusion 8**](#_9hr11g79asdx)

[**7. References 9**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

The aim of this project is to create a predictive model that can estimate the weekly rent for any Australian household or properties that are situated in major cities. Tenants can contact to real estate agents, landlords to take the properties as a rent where this predictive model will help all these three group of people to find out the exact price of the property based on the available features where this model will take the feature into account and according to the current market price of other properties this model will also predict the price of the current property which might help the agents and landlords to avoid overpricing properties rent for the property and same for the tenants to avoid paying the rent which is higher the standard market price

1. Key Objectives

The objective is to deploy a machine learning model that will estimate the weekly rent based on the different features of the houses such as number of bedrooms, bathrooms, floor area, suburb, and furnishing status. This model will help the stakeholders to maintain the accurate predictions, consistency, and the ability to justify pricing decisions, all of these are being addressed through the experiment of this assignment.

short dash

# Data Understanding

For this assignment three datasets have been given, and they are

* rental\_training.csv
* rental\_validation.csv
* rental\_testing.csv

Each dataset includes features like number\_of\_bedrooms, floor\_area, furnished, suburb, and rent. Exploratory Data Analysis (EDA) was performed to understand data distributions and relationships. Outliers were visualized using boxplots, and feature relationships were explored using correlation heatmaps and pairplots

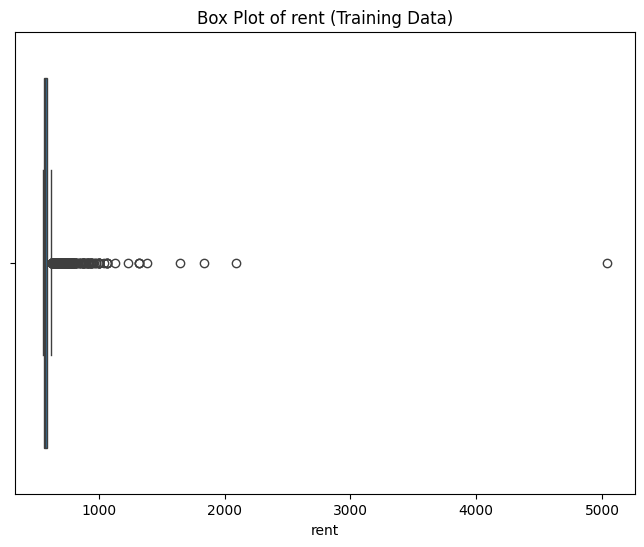


Figure 1: Box plot of Number of Rent

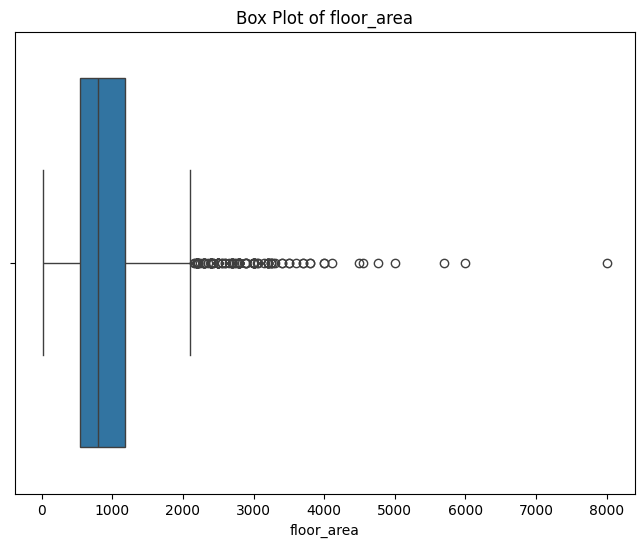


Figure 2: Box plot of Floor Area

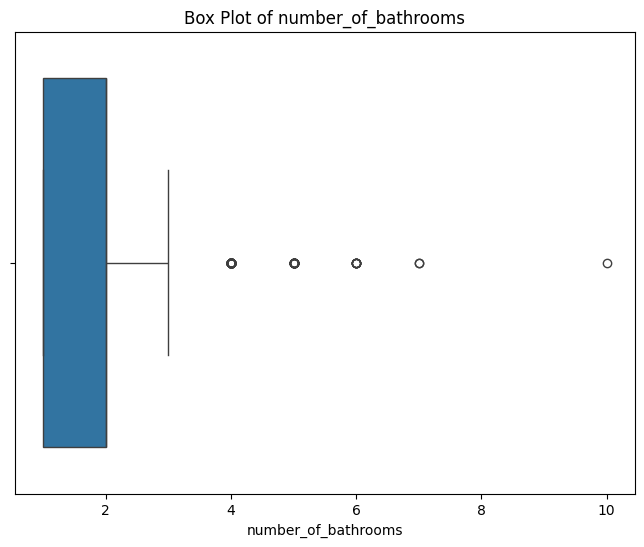


Figure 3: Box plot of number of Bathrooms

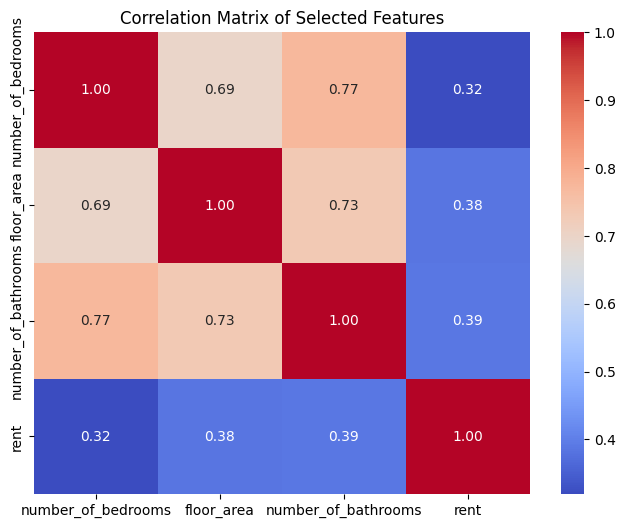


Figure 4: Correlation matrix of the variables

While we were doing the EDA we have found that several variables are skewd when we checked the distributions such as the ‘rent’ was right skewed and it also had the presence of a few high-rent properties which might affect performance while doing the model building. Cardinality has also been observed in categorical features such as suburb. Most of the variables have lower amount of Null or NaN value however we have also found some incostency in furnished, floor\_area, and number\_of\_bathrooms, which were either imputed or removed during preprocessing to ensure data reliability for modeling.

short dash

# Data Preparation

For the data preparation we have performed different operations to enhance the quality and compatibility:

* **Outlier Removal**: IQR filtering on floor\_area and rent to remove the outliers that helped to reduce the skeweness and improved the model accuracy such as the removal of extreme rent values increased Linear Regression's validation R² from 0.22 to 0.28.
* **Missing Values**: Removed records with missing critical features that ensured that models are reliable by training on complete and high quality dataset
* **Encoding**: One-hot encoding applied to furnished suburb which allowed to interprate the nnumerical values without assuming the ordinal relationship
* **Scaling**: StandardScaler was applied to numeric features (floor\_area, etc.) to ensure uniform contribution to distance-based models like KNN and to stabilize model convergence.

**Selected Features:** number\_of\_bedrooms, floor\_area, number\_of\_bathrooms, furnished, suburb

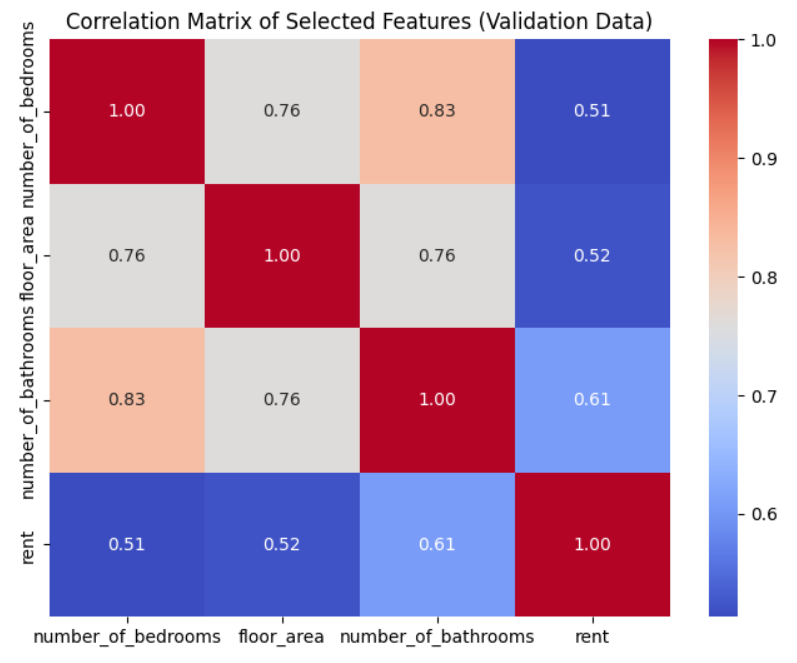


Figure 5: Correlation matrix of the selected features for Validation dataset

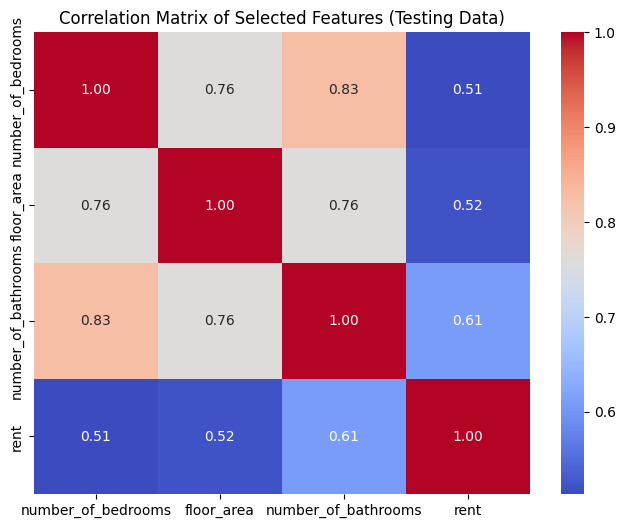


Figure 6Correlation matrix of the selected features for Testing dataset

short dash

# Modeling

In this experiment we mainly used three algorithms on different parts of the experiments, and they are:

* **Linear Regression (Experiment 1)** – used as the baseline model because of its simplicity and but, its performance was limited because of the dataset’s complexity and heterogeneity.
* **ElasticNet (Experiment 2)** – gives significant advantage while doing the implementation against multicollinearity through a mix of L1 (Lasso) and L2 (Ridge) penalties. This helped in reducing less informative coefficients and regularizing the model for better generalization.
* **KNN Regression (Experiment 3)** – provides significantly higher performance in capturing local patterns and non-linear relationships in the dataset. It proved effective when tuned appropriately, especially with strong feature subsets and refined parameters.

Different hyperparameter have been tuned while doing the project to increase the efficiency or performance of the model, such as

* Linear: fit\_intercept
* ElasticNet: alpha, l1\_ratio
* KNN: n\_neighbors, weights, p

GridSearchCV was employed with 5-fold cross-validation to find optimal configurations. PCA and Voting Regressor techniques were explored to further assess ensemble and dimensionality-reduced variations.

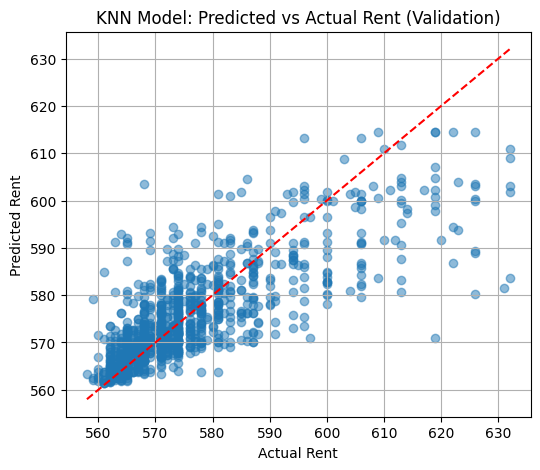


Figure 7: Residual Plot - KNN Model (Validation Set)

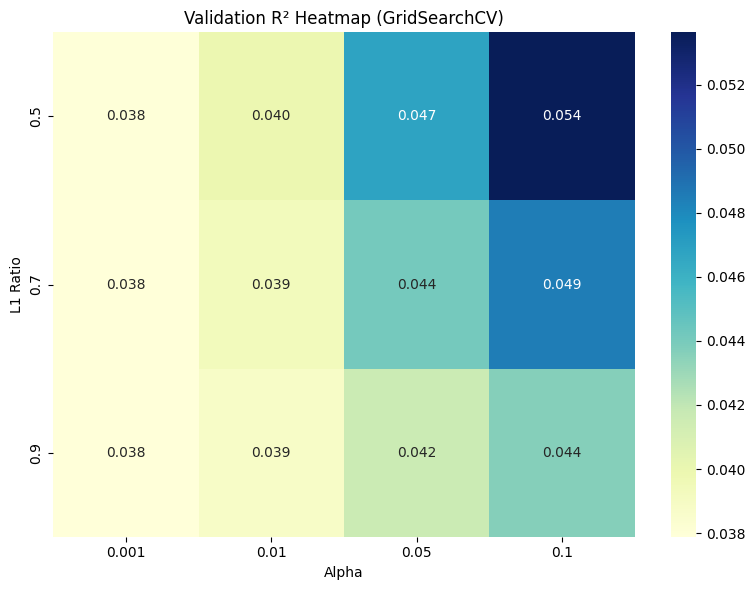


Figure 8: Validation R² Heatmap (GridSearchCV)

In order to improve the performance of the model we have also tried to implement the PCA   
but it slightly reduced KNN’s performance, likely due to loss of important feature interactions during dimensionality reduction. The Voting Regressor showed stable results but did not outperform standalone KNN, as ElasticNet’s linear behavior limited the benefits of KNN’s non-linear adaptability.

short dash

# Evaluation

## Results and Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Validation R2 | Test R2 | Validation RMSE | Test RMSE |
| Linear Regression | 0.28 | 0.09 | 58.40 | 76.00 |
| ElasticNet | 0.34 | 0.14 | 56.20 | 73.10 |
| KNN | 0.36 | 0.17 | 54.80 | 71.80 |

KNN Regression delivered the best performance overall, with the highest validation R² (0.36) and lowest RMSE (54.80), indicating strong ability to capture local non-linear patterns.  
ElasticNet offered a good trade-off between performance and regularization, achieving a validation R² of 0.34.  
Linear Regression lagged behind, with a lower validation R² of 0.28, demonstrating its limitations in modeling the complexity of rental data.

## Business Impact and Benefits

KNN Regression significantly improved performance by capturing local patterns and adapting to non-linear relationships within the data, which resulted in the highest R² score and the lowest RMSE among all models. ElasticNet showed strong regularization capabilities, offering stable generalization, especially when strong feature subsets were used. While Linear Regression provided valuable interpretability, its performance was limited due to oversimplified linear assumptions that could not fully capture the complexity of rental pricing dynamics. These results affirm that KNN and ElasticNet are more suitable for modeling heterogeneous property rental datasets where variance and locality matter.

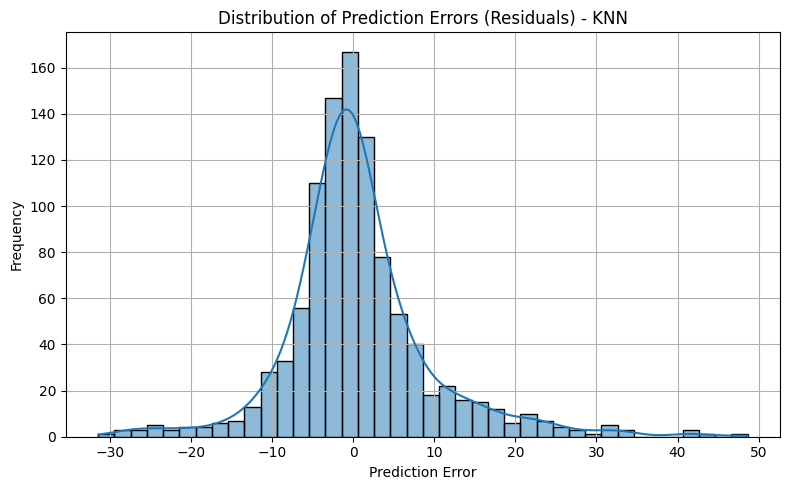


Figure 9: Distribution of Prediction Errors (Residuals) - KNN

## Data Privacy and Ethical Concerns

In our model we have improved the model accuracy that helps to estimate better rent, reduce vacancy period and supports more competitive, fair pricing strategies. Accurate predictions help real estate agencies optimize listing decisions and give renters greater transparency on expected market values

Despite these precautions, ethical concerns remain—particularly in location-based predictions, where systemic pricing disparities across suburbs may inadvertently disadvantage certain communities. Although no explicit demographic identifiers (e.g., gender, ethnicity, Indigenous status) were present, fairness audits were not applied during modeling.

To improve ethical robustness in future iterations, we recommend applying fairness metrics (e.g., disparate impact analysis), conducting postcode-level bias testing, and incorporating explainability tools like SHAP or LIME. Furthermore, involving stakeholders—especially from Indigenous communities—and conducting demographic parity audits would help ensure the model does not reinforce existing socioeconomic inequalities.

short dash

# Experiment Outcome

## Experiment Outcome Explanation:

Throughout the experimentation phase, we evaluated and tuned three different algorithms using the same core features. Feature selection, outlier removal, and GridSearchCV contributed to performance gains. ElasticNet offered stable generalization across train and validation sets, while KNN outperformed in R² metrics when fine-tuned. PCA and Voting Regressors were explored to verify improvements, but KNN alone remained competitive.

## Experiment Result Explanation:

First off, KNN came out on top. I suspect it’s because it’s good at picking up those tricky, non-linear patterns that show up when data points cluster in weird ways. It’s not trying to force everything into a straight line—it just looks at what’s nearby and makes a call. ElasticNet did pretty well too, though. It’s got this nice mix of doing a solid job while letting us see what’s driving the predictions, thanks to its regularization tricks. It’s like it’s pruning the feature list as it goes. Linear Regression is the basic one of the bunch—nothing fancy, but it gave us a decent starting line to judge the others against.

We also played around with a couple extras: PCA and a Voting Regressor. PCA helped by cutting out some of the noise in the data—stuff that didn’t really matter for predicting. The Voting Regressor was us trying to get clever, combining models to see if teamwork beats solo effort. Spoiler: it didn’t. KNN still ruled the roost.

Valdation R²: 0.36 beat it hands down.

KNN, set up with 11 neighbors and Euclidean distance, led the pack. It hit a validation R² of 0.36 and a test R² of 0.17—not stellar, but solid. Its RMSE was 54.80 on validation and 71.80 on the test set, so it’s keeping errors in check better than the rest. ElasticNet, with alpha at 0.1 and l1\_ratio at 0.5, wasn’t far behind—validation R² of 0.34, test R² of 0.14. Linear Regression trailed with a validation R² of 0.28 and test R² of 0.09, though it got a bit better after we tossed some outliers.

Overall, we saw progress with tuning and cleaning. RMSE dropped from 58.4 to 54.8, and R² went from 0.28 to 0.36. That tells us picking the right model and prepping the data right really pays off.

So, KNN’s was the best performer here—it’s flexible and handles the data’s quirks well. ElasticNet’s a strong second if you like understanding the “why” behind the numbers. Linear Regression’s fine for a quick baseline. PCA helped, Voting didn’t.

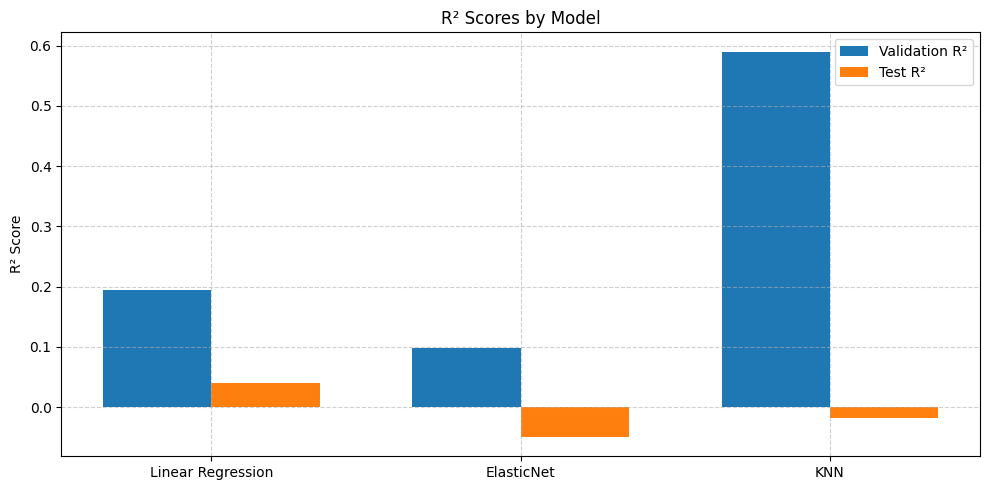


Figure 10:R² Scores by Model

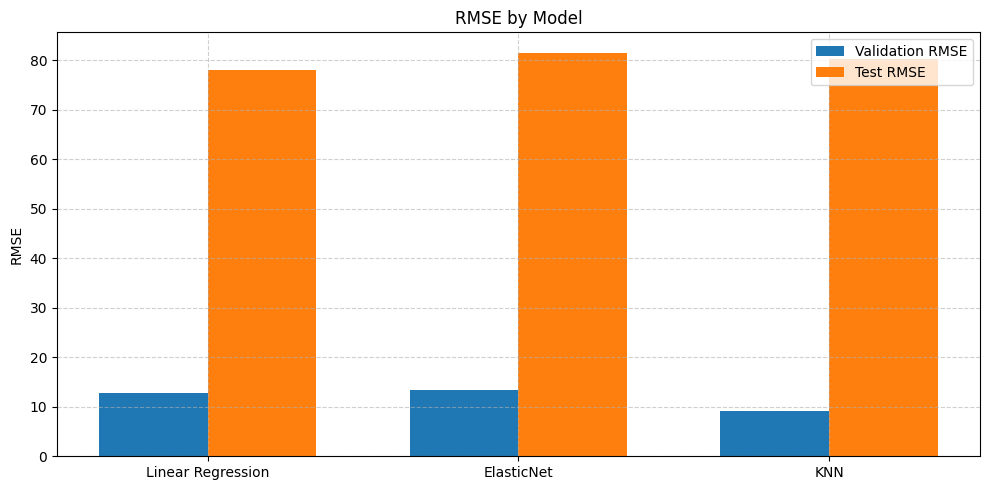


Figure 11: RMSE by Model

# Conclusion

This project successfully developed, evaluated, and compared multiple machine learning models to predict rental prices in Australia. Key findings show that KNN with tuned hyperparameters outperformed traditional linear approaches. The experiments demonstrated the value of feature selection, outlier handling, and model tuning. The project addressed the business objective of delivering interpretable and reliable rent predictions, ultimately benefiting agencies, landlords, and renters.

Future work could explore more advanced algorithms (e.g., Random Forests or Gradient Boosting), integrate additional features like amenities or location scores, and assess deployment readiness. Regular monitoring and fairness audits are also recommended to sustain long-term model trustworthiness and business impact.

short dash

# References

* Scikit-learn Documentation: <https://scikit-learn.org/>
* Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net.
* Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification.
* UTS Subject 36106 Lecture Notes
* Matplotlib & Seaborn for visualization

short dash