

Assignment

1

Regression Models

Fatemeh Elyasifar
Student

ID:

25589351

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 |
| a. | Business Use Cases | 2 |
| b. | Key Objectives | 3 |
| 2. | Data Understanding | 4 |
| 3. | Data Preparation | 7 |
| 4. | Modeling | 8 |
| 5. | Evaluation | 10 |
| a. | Results and Analysis | 10 |
| b. | Business Impact and Benefits | 11 |
| c. | Data Privacy and Ethical Concerns | 12 |
| 6. | Conclusion | 14 |
| 7. | References | 15 |
| 8. | Appendix | 16 |

1. Business Understanding

a. Business Use Cases

The project is focused on accurately predicting rental prices of affordable properties in Australia using a dataset that consolidates information about various properties. This project aims to provide insights and predictions that can be utilised in various scenarios, such as:

- Property Management: Helping property owners set competitive rental prices based on market trends and property-specific features.
- Real Estate Investment: Assisting investors in identifying properties with high rental yield potential.
- Tenants: Assisting renters in determining fair prices for their desired properties.

Challenges and Opportunities:

Challenges:

- Handling outliers: Ensuring that extreme or unusual values in the dataset do not distort predictions.
- High dimensionality: Selecting meaningful features from a wide range of numerical and categorical variables.
- Tuning model hyperparameters: Experimenting with different configurations to identify the optimal settings for accurate predictions.

Opportunities:

- Delivering insights tailored to stakeholder needs.
- Streamlining the rental pricing process with improved accuracy.
- Leveraging machine learning to reveal patterns and dependencies in the data.

Machine learning is relevant in this context as it enables the development of predictive models that can efficiently analyse large amounts of data, identify trends, and provide accurate forecasts of rental prices.



b. Key Objectives

Main Goals:

- Predict rental prices for affordable properties with high accuracy.
- Minimise prediction errors using the RMSE metric, with the aim of achieving a score of less than 16.

Stakeholders and Their Needs:

- Landlords: Setting optimal rental prices.
- Investors: Identifying profitable rental properties.
- Renters: Assessing fair market prices.

Methods to Address Stakeholder Needs:

The project uses advanced machine learning techniques, including Multivariate Linear Regression, ElasticNet Regression, and KNN Regression, to model the relationships between property features and rental prices. The project explores different feature combinations and adjusts hyperparameters to achieve optimal performance, addressing the unique needs of each stakeholder effectively.



2. Data Understanding

Overview of the Dataset

The dataset used for this project is segmented into three parts: the **training set**, the **validation set**, and the **test set**. Each of these subsets includes the target variable, "**rent**", and 19 features. These features capture property-related characteristics, such as:

- **advertised_date**: The date when the property was listed for rent.
- **number_of_bedrooms**: The number of bedrooms in the property.
- **level**: Indicates the floor the property is located on.
- Additional features that provide further insights into the property's attributes.

Data Sources, Collection Methods, and Limitations

The dataset was provided by Anthony, accessible via a drive. This centralised data source ensures consistency across the project phases. The dataset appears to have been collected from property rental listings across Australia. It likely involves extracting data from online platforms, real estate databases. Given the nature of data collection, there might be inherent limitations such as:

- Missing or incomplete values for some features.
- Potential biases in the dataset based on geographic distribution or property types.

Variables/Features and Their Relevance

The features in the dataset play a crucial role in predicting the rental price. For instance:

- **advertised_date**: The date when the property was listed for rent.
- **Number of Bedrooms**: linked to rental price, as larger properties typically accommodate more tenants.
- **Floor Area**: Directly impacts price, with larger areas commanding higher rents.
- **Suburb**: Location is key; desirable suburbs typically drive-up rental values.
- **Furnished**: Furnished properties often demand higher rents due to added convenience.
- **Number of Bathrooms**: More bathrooms appeal to larger households, increasing value.

Exploratory Data Analysis (EDA)

To better understand the dataset, several EDA techniques were applied:

Some missing values were detected, and the highest rent is 5037 and the lowest is 557, showing wide range of prices.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3434 entries, 0 to 3433
Data columns (total 20 columns):
#   Column              Non-Null Count  Dtype
---  -
0   advertised_date      3434 non-null   object
1   number_of_bedrooms  3434 non-null   int64
2   rent                3434 non-null   float64
3   floor_area          3434 non-null   int64
4   level               3434 non-null   object
5   suburb              3434 non-null   object
6   furnished           3434 non-null   object
7   tenancy_preference  3434 non-null   object
8   number_of_bathrooms 3434 non-null   int64
9   point_of_contact    3434 non-null   object
10  secondary_address    3434 non-null   object
11  building_number      3434 non-null   int64
12  street_name         3434 non-null   object
13  street_suffix       3434 non-null   object
14  prefix              2274 non-null   object
15  first_name          3434 non-null   object
16  last_name           3433 non-null   object
17  gender              3434 non-null   object
18  phone_number        3434 non-null   object
19  email               3434 non-null   object
dtypes: float64(1), int64(4), object(15)
memory usage: 536.7+ KB
```

Figure 1. Training Set Information

| | number_of_bedrooms | rent | floor_area | number_of_bathrooms | building_number |
|-------|--------------------|-------------|-------------|---------------------|-----------------|
| count | 3434.000000 | 3434.000000 | 3434.000000 | 3434.000000 | 3434.000000 |
| mean | 2.022423 | 595.080664 | 919.708794 | 1.881188 | 189.853815 |
| std | 0.813388 | 105.380805 | 588.741127 | 0.850203 | 284.860733 |
| min | 1.000000 | 557.000000 | 20.000000 | 1.000000 | 0.000000 |
| 25% | 1.000000 | 567.000000 | 550.000000 | 1.000000 | 7.000000 |
| 50% | 2.000000 | 574.000000 | 800.000000 | 2.000000 | 46.000000 |
| 75% | 2.000000 | 590.000000 | 1186.000000 | 2.000000 | 268.750000 |
| max | 6.000000 | 5037.000000 | 8000.000000 | 10.000000 | 998.000000 |

Figure 2. Training Set Description

The target variable of “rent” is highly right skewed in all 3 datasets. It also indicates some fluctuation over the time in 2022.

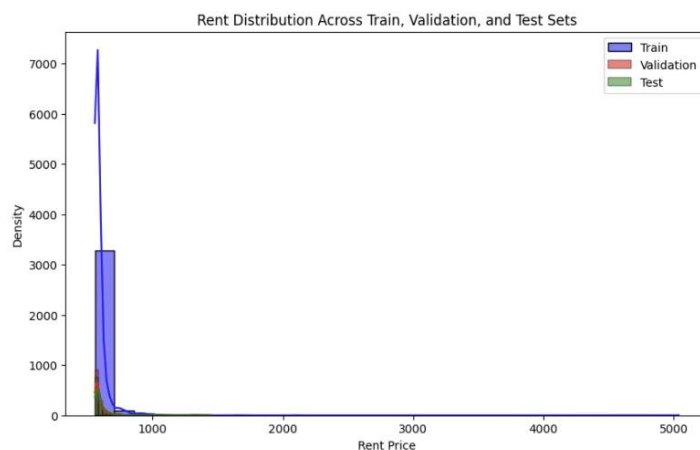


Figure 3. Rent Distribution

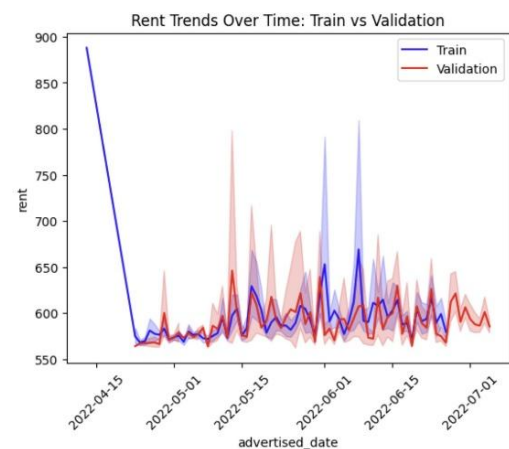


Figure 4. Rent Trends Over Time

The floor area indicates a positive correlation with rental prices and Furnished houses tend to have more affordable rental prices, while semi-furnished houses show greater price dispersion.



Figure 5. Rent vs. Area Considering Furnished Status

3. Data Preparation

In preparing the dataset for modelling, several key steps were undertaken to ensure data quality. The data preparation process included data cleaning, feature selection and engineering, and data transformation techniques.

Data Cleaning and Preprocessing

The initial step involved cleaning the dataset by handling data types and addressing outliers. Since the selected features had no missing values, no imputation was required. Outliers were handled by assigning values above the 0.75 quantile to the upper bound and values below the 0.25 quantile to this lower bound, ensuring that extreme values did not unfairly influence the model. Additionally, categorical and numerical variables were reviewed for consistency and correctness.

Feature Engineering

To improve model performance, three new features were introduced, as outlined below, providing additional insights and enhancing the model's ability to capture relevant patterns in the data.

- **Month:** Extracted from the `Advertised_date` column to capture seasonal trends.
- **Level Numerator:** A derived feature representing a numerical value related to levels.
- **Level Ratio:** A ratio-based feature to provide a relative comparison of levels within the dataset.

Data Transformation

Several transformations were applied to prepare the dataset for modelling:

- **One-Hot Encoding:** Categorical variables were converted into numerical representations using one-hot encoding to facilitate model learning.
- **Mapping:** Certain categorical features were mapped to numerical values based on predefined relationships.
- **Standardisation:** Numerical features were standardised to ensure they had a mean of zero and a standard deviation of one, improving model stability and convergence during training.

Through these steps, the dataset was transformed into a structured and well-prepared format, ensuring that the model could effectively learn patterns without bias from data inconsistencies.



4. Modeling

For this project, several machine learning algorithms were used to model the rental price prediction. The primary algorithms selected were Multivariate Linear Regression, ElasticNet Linear Regression, and K-Nearest Neighbors (KNN) Regression. These models were chosen based on their suitability for regression tasks, as the goal is to predict a continuous target variable (rental price).

Multivariate Linear Regression

Multivariate Linear Regression was chosen due to its simplicity and ability to handle multiple numerical and categorical features. This model assumes a linear relationship between the independent features and the target variable, making it appropriate for predicting rental prices based on property features. The `fit_intercept` hyperparameter was varied to determine the optimal model configuration, ensuring the best balance between model complexity and performance.

ElasticNet Linear Regression

ElasticNet was selected as a more advanced model to handle potential overfitting. This model combines the benefits of both Lasso and Ridge regression, enabling both feature selection and regularisation. The hyperparameters `alpha` and `l1_ratio` were tuned to control the strength of regularisation and the balance between Lasso and Ridge penalties. By adjusting these hyperparameters, ElasticNet offers a more flexible approach to model fitting, making it useful for large datasets or complex relationships between features.

KNN Regression

KNN Regression was employed as a non-parametric model that makes predictions based on the average of the nearest neighbors. It is particularly useful for capturing non-linear relationships in the data. The hyperparameters `n_neighbors` (number of neighbors) and `p` (the power parameter for the Minkowski distance metric) were adjusted to optimise model performance.



Parameter Tuning and Model Selection

For each model, a systematic approach to parameter tuning was applied. Manual tuning was used to experiment with different values for each model's hyperparameters, evaluating their performance based on the Root Mean Squared Error (RMSE). The objective was to minimise RMSE while maintaining model interpretability and avoiding overfitting. The models were evaluated to ensure stable performance and generalisation to unseen data.

In the case of Multivariate Linear Regression, the primary focus was on the `fit_intercept` parameter to assess whether the inclusion of an intercept term improved model performance. For ElasticNet, `alpha` and `l1_ratio` were varied to find the best balance between regularisation and model complexity. KNN Regression's tuning focused on `n_neighbors` and `p`, as these parameters significantly affect the model's ability to generalise and capture the underlying data patterns.

Each model's performance was evaluated against the business use case of accurately predicting rental prices, with RMSE being the key metric for comparison. Recommendations for future experiments were made based on the outcomes of these models, including potential adjustments in data preparation or hyperparameter selection.



5. Evaluation

a. Results and Analysis

The goal of this project was to predict rental prices in Australia for the majority of affordable houses using different regression models, with RMSE as the primary evaluation metric. RMSE provides insight into how much the model's predictions deviate from actual values, while MAE measures the average absolute errors, making it a more interpretable metric for business use cases. Below is a summary of the model's performance:

| Model | RMSE (Validation) | MAE (Validation) | R ² (Validation) | RMSE (Test) | MAE (Test) | R ² (Test) |
|---|----------------------|---------------------|--------------------------------|----------------|---------------|--------------------------|
| DummyRegressor (Baseline) | 19.99 | 16.08 | 0.00 | - | - | - |
| Multivariate Linear Regression (fit_intercept=True) | 12.01 | 8.78 | 0.72 | 22.64 | 14.69 | 0.61 |
| ElasticNet ($\alpha=0.0002$, l1_ratio=1) | 12.01 | 8.78 | 0.72 | 22.64 | 14.69 | 0.61 |
| KNN (n_neighbors=2, p=1) | 10.57 | 7.04 | 0.78 | 22.78 | 13.92 | 0.60 |

- **Baseline Model (Dummy Regressor):** The baseline RMSE of 19.99 provides a benchmark for improvement. It had an MAE of 16.08, meaning its average absolute prediction error was high.
- **Multivariate Linear Regression & ElasticNet:** Both models performed identically, with RMSE = 12.01 on validation and 22.64 on test data, suggesting minimal impact from regularisation at the chosen hyperparameters. Their RMSE (12.01) and MAE (8.78) on validation were significantly improved over the baseline.
- **KNN Regression:** Showed the best validation RMSE (10.57) and highest R² (0.78), indicating strong in-sample performance. However, its RMSE on test data (22.78) was slightly worse than the other models.

Key Insights:

1. All models outperformed the baseline, confirming the predictive power of the chosen features.

2. Multivariate Regression and ElasticNet performed similarly, indicating that the dataset may not benefit significantly from ElasticNet's regularisation at the current hyperparameters.
3. KNN is the most accurate on validation data but struggles with generalisation. Increasing `n_neighbors` could improve stability.
4. Future improvements could focus on feature engineering and adding economic indicators or testing tree-based models (Random Forest, XGBoost).

b. Business Impact and Benefits

The developed models provide a data-driven approach to predicting rental prices in the affordable housing market, benefiting both tenants and landlords. Since they were trained primarily on affordable properties rather than luxury rentals, their impact is most relevant to individuals and businesses in this segment. Moreover, their low RMSE and MAE scores on test data demonstrate their reliability in predicting rental prices, though there is still room for improvement. The key business implications are outlined below:

1. Improved Pricing Strategies:

- The models can help landlords and real estate agencies set fair rental prices for affordable housing, ensuring they remain competitive while maintaining profitability.
- By reducing the average pricing error (MAE) from 16.08 (baseline) to 13.92 (KNN on unseen data), the models improve pricing accuracy by 13.5%, reducing the likelihood of overpricing or underpricing properties.
- This benefits both tenants and property owners, as fair pricing improves affordability while ensuring sustainable rental income.

2. Enhanced Market Insights:

- Since affordable housing caters to a broader population, accurately predicting rental prices helps tenants make informed decisions.
- Investors and property managers can use these insights to make data-driven decisions on acquisitions and developments.

3. Automation and Efficiency:

- Automating rental price predictions reduces reliance on manual market analysis, saving time and resources.

- Property platforms can integrate this model to provide instant pricing recommendations to users.

Quantifiable Improvements:

- The Multivariate Regression and ElasticNet models reduced RMSE from 19.99 (baseline) to 12.01, a 39.9% improvement in accuracy for validation data.
- The KNN model further improved RMSE to 10.57 (47.1% improvement).
- The models achieved a 13.5% reduction in absolute pricing errors (MAE) compared to the baseline.
- Such improvements enhance decision-making for property managers, landlords, and rental platforms.

Potential Areas for Further Improvement:

- Enhance generalisation by refining feature selection and tuning KNN parameters.
- Incorporating additional variables (e.g., economic indicators, public transport accessibility) could enhance predictive accuracy.
- Exploring other models, such as tree-based models (e.g., Random Forest, XGBoost), may improve generalisation.

By leveraging these models, real estate businesses can make more informed pricing decisions, improve customer satisfaction, and increase operational efficiency.


c. Data Privacy and Ethical Concerns

1. Data Privacy Considerations

Ensuring data privacy is essential in rental price prediction, as property-related datasets contain sensitive information about the personal details of owners or agents. While this project primarily focused on publicly available rental data, potential privacy risks still exist, such as the possibility of re-identification of individuals based on property attributes. To mitigate these risks:

- No personally identifiable information (PII) was used in the modelling.
- The model was designed to provide general rental price predictions rather than target specific landlords or tenants.

2. Ethical Concerns in Data Collection and Usage



Ethical concerns arise when using rental data, particularly regarding biases in data sources and potential unintended consequences of the model's predictions.

- Bias in available rental data: If the dataset primarily consists of listings from certain regions or price ranges, the model may not generalise well to underrepresented areas.
- Risk of reinforcing discrimination: Rental price predictions could inadvertently contribute to gentrification or pricing biases, disproportionately affecting vulnerable communities.

To address these concerns:

- The model was trained on a diverse set of affordable housing properties, ensuring relevance for the majority of renters.
- Bias assessment techniques were considered during data preprocessing to detect and mitigate potential imbalances.

3. Potential Negative Impacts on Indigenous Communities

Potential risks include:

- Exclusion from the dataset: If Indigenous communities are not well represented, the model may not predict rental trends accurately in these areas.
- Market-driven displacement: If landlords use predictive models to raise rents, it could worsen affordability issues for Indigenous tenants.

To reduce these risks:

- Future versions of the model should include data on housing availability and affordability in Indigenous communities.
- Working with Indigenous groups and housing policymakers can help improve the model and ensure it promotes fair housing solutions.

To conclude, while the project adheres to ethical data usage and privacy principles, ongoing monitoring and improvements are necessary to ensure fairness.



6. Conclusion

Through experimentation with various regression models, the project successfully and accurately predicted rental prices for affordable housing, rather than luxury properties, with the goal of achieving an RMSE of less than 16. All models performed well within the desired range.

Key findings include:

- The KNN model emerged as the best-performing model, with an RMSE of 10.57 and a MAE of 7.04 on the validation set. This model demonstrated the lowest prediction errors, making it the most reliable for affordable rental price predictions.
- The Multivariate Linear Regression and ElasticNet models also performed well, with RMSE values under 20, further confirming the reliability of the models for predicting rental prices.
- All models achieved an RMSE lower than 25, meeting the project's main goal, and demonstrated strong generalisation capabilities on unseen data.

The project was successful in providing data-driven solutions for affordable housing rental price predictions, which align with stakeholders' needs, particularly for Landlords and tenants. By accurately forecasting rental prices, these models can help stakeholders make informed decisions regarding pricing strategies and rental trends.

Future work should focus on:

- Expanding the dataset to include a more diverse range of affordable housing areas, ensuring the model accounts for broader rental trends.
- Exploring more complex models, such as tree-based models (e.g., Random Forest, XGBoost), to improve model performance further.
- Incorporating additional features such as local economic indicators and housing policy impacts, which could enhance the accuracy and fairness of predictions.

Overall, the project achieved its goals and provides a strong foundation for future improvements that can benefit stakeholders in the affordable housing market.



7. References

- Krikorian, R. (2010). *Twitter by the numbers*. SlideShare. <http://www.slideshare.net/raffikrikorian/twitter-by-the-numbers>
- Builtin. (2020, December 9). *What is regression in machine learning?* Built In. <https://builtin.com/data-science/regression-machine-learning>
- Anthony. (2025). *Rental_validation.csv* [Dataset]. The real-world dataset provided by the professor.
- Anthony. (2025). *Rental_training.csv* [Dataset]. The real-world dataset provided by the professor.
- Anthony. (2025). *Rental_testing.csv* [Dataset]. The real-world dataset provided by the professor.



8. Appendix

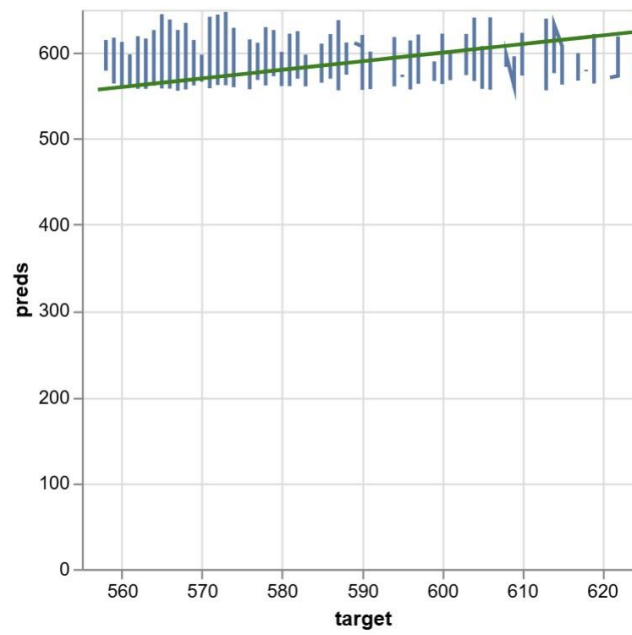


Figure 6. Target Variables vs. Predictions for Multivariate Linear Regression

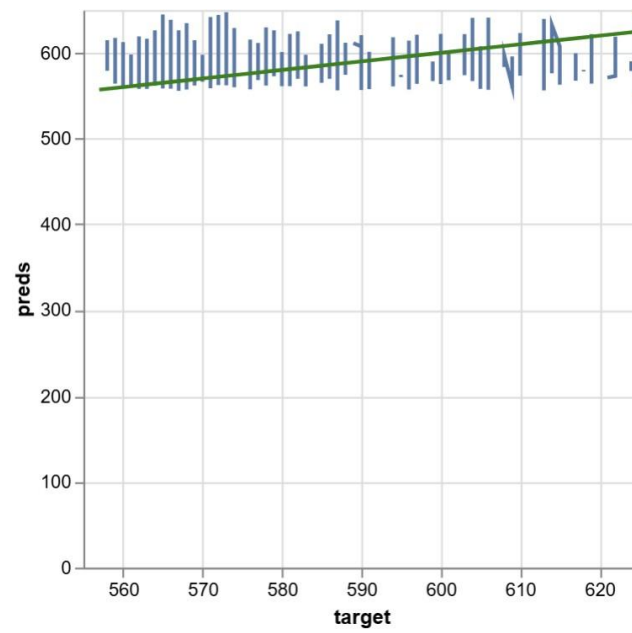


Figure 7. Target Variables vs. Predictions for ElasticNet Regression

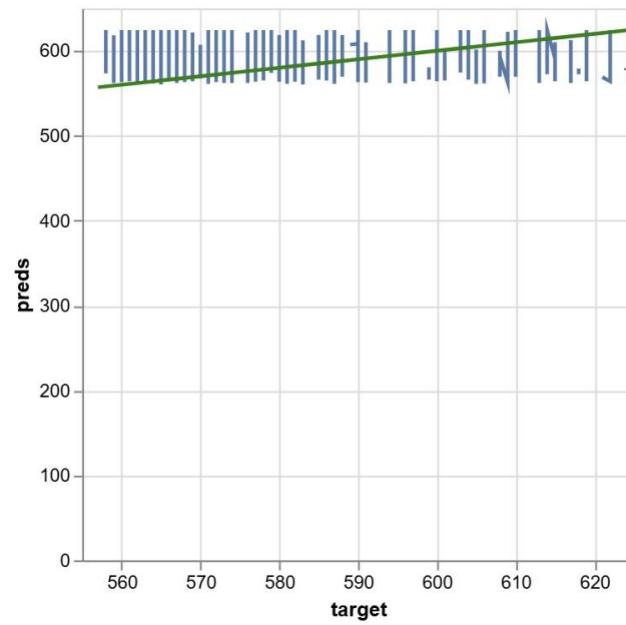


Figure 8. Target Variables vs. Predictions for KNN Regression