# 36106\_25AU-AT1\_25589351\_experiment\_1

March 29, 2025

## 1 Experiment Notebook

1.1 0. Setup Environment

## 1.1.1 0.a Install Mandatory Packages

Do not modify this code before running it

```
[1]: # Do not modify this code
     import os
     import sys
     from pathlib import Path
     COURSE = "36106"
     ASSIGNMENT = "AT1"
     DATA = "data"
     asgmt_path = f"{COURSE}/assignment/{ASSIGNMENT}"
     root_path = "./"
     print("##### Install required Python packages #####")
     | pip install -r https://raw.githubusercontent.com/aso-uts/labs_datasets/main/
      →36106-mlaa/requirements.txt
     if os.getenv("COLAB_RELEASE_TAG"):
        from google.colab import drive
        from pathlib import Path
        print("\n##### Connect to personal Google Drive #####")
        gdrive_path = "/content/gdrive"
        drive.mount(gdrive_path)
        root_path = f"{gdrive_path}/MyDrive/"
     print("\n##### Setting up folders #####")
     folder_path = Path(f"{root_path}/{asgmt_path}/") / DATA
```

```
folder_path.mkdir(parents=True, exist_ok=True)
print(f"\nYou can now save your data files in: {folder_path}")
if os.getenv("COLAB_RELEASE_TAG"):
    %cd {folder_path}
###### Install required Python packages ######
Requirement already satisfied: pandas==2.2.2 in /usr/local/lib/python3.11/dist-
packages (from -r https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 1)) (2.2.2)
Requirement already satisfied: scikit-learn==1.6.1 in
/usr/local/lib/python3.11/dist-packages (from -r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 2)) (1.6.1)
Requirement already satisfied: altair==5.5.0 in /usr/local/lib/python3.11/dist-
packages (from -r https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (5.5.0)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-
packages (from pandas==2.2.2->-r https://raw.githubusercontent.com/aso-
```

```
Requirement already satisfied: narwhals>=1.14.2 in
/usr/local/lib/python3.11/dist-packages (from altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (1.31.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from altair==5.5.0->-r https://raw.githubusercontent.com/aso-
uts/labs datasets/main/36106-mlaa/requirements.txt (line 3)) (24.2)
Requirement already satisfied: typing-extensions>=4.10.0 in
/usr/local/lib/python3.11/dist-packages (from altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (4.12.2)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.11/dist-
packages (from jsonschema>=3.0->altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (25.3.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=3.0->altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.11/dist-packages (from jsonschema>=3.0->altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-
packages (from jsonschema>=3.0->altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 3)) (0.23.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas==2.2.2->-r
https://raw.githubusercontent.com/aso-
uts/labs_datasets/main/36106-mlaa/requirements.txt (line 1)) (1.17.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->altair==5.5.0->-r
https://raw.githubusercontent.com/aso-
uts/labs datasets/main/36106-mlaa/requirements.txt (line 3)) (3.0.2)
###### Connect to personal Google Drive ######
Mounted at /content/gdrive
##### Setting up folders #####
You can now save your data files in:
/content/gdrive/MyDrive/36106/assignment/AT1/data
/content/gdrive/MyDrive/36106/assignment/AT1/data
```

## 1.1.2 0.b Disable Warnings Messages

Do not modify this code before running it

```
[]: import warnings warnings.simplefilter(action='ignore', category=FutureWarning)
```

#### 1.1.3 0.c Install Additional Packages

If you are using additional packages, you need to install them here using the command: ! pip install package\_name>

```
[]: # <Student to fill this section>
```

#### 1.1.4 0.d Import Packages

```
[2]: import ipywidgets as widgets
import pandas as pd
import altair as alt
import numpy as np
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
```

## 1.2 A. Project Description

HBox(children=(Text(value='Fatemeh Elyasifar', description='Student Name:', placeholder='<student to fill this...

```
[4]: print("Student Name:", wgt_student_name.value)
print("Student Id:", wgt_student_id.value)
```

Student Name: Fatemeh Elyasifar Student Id: 25589351

```
[5]: # @title Experiment ID

wgt_experiment_id = widgets.BoundedIntText(
    value="1",
    min=0,
    max=3,
    step=1,
    description='Experiment ID:',
    style={'description_width': 'initial'},
    disabled=False
)
wgt_experiment_id
```

BoundedIntText(value=1, description='Experiment ID:', max=3, style=DescriptionStyle(description\_width='initial...

```
[6]: print("Experiment ID:", wgt_experiment_id.value)
```

Experiment ID: 1

```
[7]: # @title Business Objective
     wgt_business_objective = widgets.Textarea(
          value="The main objective is to develop a machine learning model that
       \hookrightarrowaccurately predicts rental prices specifically for affordable properties in\sqcup
       \hookrightarrowAustralia, excluding luxury homes. This model aims to assist real estate\sqcup
       \hookrightarrowagencies, property investors, and tenants in making informed decisions based_{\sqcup}
       \hookrightarrowon the features of affordable housing and market trends. The key success_{\sqcup}
       \hookrightarrowmetric is RMSE, with the goal of achieving an RMSE score of less than 16 on_{\sqcup}
       \hookrightarrowthe validation set, which quantifies the prediction error. Additionally,\sqcup
       \hookrightarrow feature tuning will be necessary to optimise the model's performance and
       \hookrightarrowensure accurate predictions based on the most relevant property\sqcup
       ⇔characteristics.",
          placeholder='<student to fill this section>',
          description='Business Objective:',
          disabled=False,
          style={'description width': 'initial'},
          layout=widgets.Layout(height="100%", width="auto")
     wgt_business_objective
```

Textarea(value='The main objective is to develop a machine learning model that ⊔ →accurately predicts rental pric...

```
[8]: print("Business Objective:", wgt_business_objective.value)
```

Business Objective: The main objective is to develop a machine learning model that accurately predicts rental prices specifically for affordable properties in Australia, excluding luxury homes. This model aims to assist real estate agencies, property investors, and tenants in making informed decisions based on the features of affordable housing and market trends. The key success metric is RMSE, with the goal of achieving an RMSE score of less than 16 on the validation set, which quantifies the prediction error. Additionally, feature tuning will be necessary to optimise the model's performance and ensure accurate predictions based on the most relevant property characteristics.

## 1.3 B. Experiment Description

Textarea(value='Including the intercept term in the Multivariate Linear →Regression model will improve its pred...

```
[10]: print("Experiment Hypothesis:", wgt_experiment_hypothesis.value)
```

Experiment Hypothesis: Including the intercept term in the Multivariate Linear Regression model will improve its predictive performance, as measured by RMSE and  $R^2$ , compared to excluding the intercept term. Null Hypothesis (H): Changing the fit\_intercept parameter does not significantly affect the RMSE of the Multivariate Linear Regression model.

```
placeholder='<student to fill this section>',
  description='Experiment Expectations:',
  disabled=False,
  style={'description_width': 'initial'},
  layout=widgets.Layout(height="100%", width="auto")
)
wgt_experiment_expectations
```

Textarea(value='Train multiple models with different fit\_intercept settings. 

→Compare RMSE values to evaluate t...

```
[12]: print("Experiment Expectations:", wgt_experiment_expectations.value)
```

Experiment Expectations: Train multiple models with different fit\_intercept settings. Compare RMSE values to evaluate the impact of this hyperparameter. Provide insights on whether the dataset benefits from including an intercept. Recommend potential data preprocessing improvements or feature transformations.

## 1.4 C. Data Understanding

#### 1.4.1 C.1 Load Datasets

Do not change this code

```
[3]: # Load training data
    X_train = pd.read_csv(folder_path / 'X_train.csv')
    y_train = pd.read_csv(folder_path / 'y_train.csv')

[4]: # Load validation data
    X_val = pd.read_csv(folder_path / 'X_val.csv')
    y_val = pd.read_csv(folder_path / 'y_val.csv')

[5]: # Load testing data
    X_test = pd.read_csv(folder_path / 'X_test.csv')
    y_test = pd.read_csv(folder_path / 'y_test.csv')
```

#### 1.5 D. Feature Selection

```
[]: # <Student to fill this section>

features_list = □

□ □ ['number_of_bedrooms', 'floor_area', 'number_of_bathrooms', 'month', 'level_numerator', 'level_r

□ Family', 'tenancy_preference_Family', 'suburb_Brisbane', 'suburb_Canberra', 'suburb_Melbourne',
```

```
wgt_feat_selection_explanation = widgets.Textarea(
    value="These features provide a comprehensive view of the property's_
    characteristics and market context, and were chosen to capture the key_
    factors that influence rent prices. They can help improve the accuracy of_
    rent price predictions.",
    placeholder='<student to fill this section>',
    description='Feature Selection Explanation:',
    disabled=False,
    style={'description_width': 'initial'},
    layout=widgets.Layout(height="100%", width="auto")
)
wgt_feat_selection_explanation
```

Textarea(value='These features provide a comprehensive view of the property's

substantial contents of the property's contents of the property of

[14]: print("Feature Selection Explanation:", wgt\_feat\_selection\_explanation.value)

Feature Selection Explanation: These features provide a comprehensive view of the property's characteristics and market context, and were chosen to capture the key factors that influence rent prices. They can help improve the accuracy of rent price predictions.

## 1.6 E. Train Machine Learning Model

#### 1.6.1 E.1 Import Algorithm

Provide some explanations on why you believe this algorithm is a good fit

```
[6]:  # <Student to fill this section>
from sklearn.linear_model import LinearRegression
```

```
wgt_algo_selection_explanation = widgets.Textarea(
    value="LinearRegression was selected because it effectively models the
    ⇒relationship between multiple independent variables and a dependent variable.
    ↓ It assumes a linear relationship, making it both interpretable and
    ⇒computationally efficient. This model enables comparison with the baseline
    ⇒model to assess improvements in predictive performance.",
    placeholder='<student to fill this section>',
    description='Algorithm Selection Explanation:',
    disabled=False,
    style={'description_width': 'initial'},
    layout=widgets.Layout(height="100%", width="auto")
```

```
wgt_algo_selection_explanation
```

Textarea(value='LinearRegression was selected because it effectively models the →relationship between multiple ...

```
[16]: print("Algorithm Selection Explanation:", wgt_algo_selection_explanation.value)
```

Algorithm Selection Explanation: LinearRegression was selected because it effectively models the relationship between multiple independent variables and a dependent variable. It assumes a linear relationship, making it both interpretable and computationally efficient. This model enables comparison with the baseline model to assess improvements in predictive performance.

#### 1.6.2 E.2 Set Hyperparameters

Provide some explanations on why you believe this algorithm is a good fit

```
[7]:  # <Student to fill this section>
fit_intercept_values = [True, False]
```

```
wgt_hyperparams_selection_explanation = widgets.Textarea(
    value="The fit_intercept hyperparameter was selected to evaluate whether
    including an intercept improves model performance. When fit_intercept=True,
    the model learns an intercept term, representing the baseline rent when all
    features are zero. When fit_intercept=False, the model assumes the data is
    already centered. Comparing both settings helps identify the best approach
    for accurate rent prediction.",
    placeholder='<student to fill this section>',
    description='Hyperparameters Selection Explanation:',
    disabled=False,
    style={'description_width': 'initial'},
    layout=widgets.Layout(height="100%", width="auto")
)
wgt_hyperparams_selection_explanation
```

Textarea(value='The fit\_intercept hyperparameter was selected to evaluate → whether including an intercept impro...

Hyperparameters Selection Explanation: The fit\_intercept hyperparameter was selected to evaluate whether including an intercept improves model performance. When fit\_intercept=True, the model learns an intercept term, representing the baseline rent when all features are zero. When fit\_intercept=False, the model

assumes the data is already centered. Comparing both settings helps identify the best approach for accurate rent prediction.

#### 1.6.3 E.3 Fit Model

#### 1.6.4 E.4 Model Technical Performance

Provide some explanations on model performance

```
[10]: mse_with_intercept = mean_squared_error(y_val, y_preds_with_intercept)
     rmse with intercept = np.sqrt(mse with intercept)
     mae_with_intercept = mean_absolute_error(y_val, y_preds_with_intercept)
     r2_with_intercept = r2_score(y_val, y_preds_with_intercept)
     mse_without_intercept = mean_squared_error(y_val, y_preds_without_intercept)
     rmse_without_intercept = np.sqrt(mse_without_intercept)
     mae_without_intercept = mean_absolute_error(y_val, y_preds_without_intercept)
     r2_without_intercept = r2_score(y_val, y_preds_without_intercept)
     print(f"Fit Intercept = {fit_intercept_values[0]}:")
     print(f"RMSE: {rmse_with_intercept}")
     print(f"MAE: {mae_with_intercept}")
     print(f"R2 Score: {r2 with intercept}")
     print("....")
     print(f"Fit Intercept = {fit_intercept_values[1]}:")
     print(f"RMSE: {rmse without intercept}")
     print(f"MAE: {mae_without_intercept}")
     print(f"R2 Score: {r2_without_intercept}")
      (y_preds_with_intercept - y_val).sum()
      (y_preds_without_intercept - y_val).sum()
```

Fit Intercept = True: RMSE: 12.00599255091588 MAE: 8.775523306284954 R<sup>2</sup> Score: 0.7212732502406147 Fit Intercept = False:

RMSE: 583.0052384027257

MAE: 582.8825069318831

R<sup>2</sup> Score: -656.2452034925926

[10]: rent -769404.90915

dtype: float64

Textarea(value="With the intercept included (Fit Intercept = True), the model  $_{\square}$   $_{\square}$  achieved an RMSE of 12 and an  $\mathbb{R}^2$ ...

```
[20]: print("Model Performance Explanation:", wgt_model_performance_explanation.value)
```

Model Performance Explanation: With the intercept included (Fit Intercept = True), the model achieved an RMSE of 12 and an  $R^2$  score of 0.72, indicating good performance in capturing rent price variance. However, when the intercept was excluded (Fit Intercept = False), the model's performance drastically worsened, with an RMSE of 583 and a negative  $R^2$  score of -656.25, suggesting that removing the intercept led to a poor fit and a model that failed to capture the underlying patterns in the data.

#### 1.6.5 E.5 Business Impact from Current Model Performance

Provide some analysis on the model impacts from the business point of view

```
[]: # <Student to fill this section>

y_1 = y_train['rent']
y_pred = pd.DataFrame(y_preds_with_intercept, columns=['rent_pred'])
```

## []: alt.LayerChart(...)

```
[11]: model_final = LinearRegression(fit_intercept=True)
    model_final.fit(X_train, y_train)

y_pred_final = model_final.predict(X_test)

mse_final = mean_squared_error(y_test, y_pred_final)
    rmse_final = np.sqrt(mse_final)
    mae_final = mean_absolute_error(y_test, y_pred_final)
    r2_final = r2_score(y_test, y_pred_final)

print(f"RMSE: {rmse_final}")
    print(f"MAE: {mae_final}")
    print(f"R2: {r2_final}")
```

RMSE: 22.643786463053928 MAE: 14.68910446492067 R2: 0.6113806034115141

```
wgt_model_business_explanation = widgets.Textarea(
    value="The model provides a moderate estimate of rental prices, with an R²_
    score of 0.61, capturing a significant portion of rent price variance._
    However, with an RMSE of 12 on the training set and 22.64 on unseen data,_
    there is room for improvement. The higher RMSE on unseen data suggests that_
    the model may not generalise well, leading to occasional mispricing that_
    could impact revenue and tenant satisfaction. To improve performance,_
    exploring regularised models like Ridge or Lasso could help reduce errors_
    and enhance pricing accuracy.",
    placeholder='<student to fill this section>',
```

```
description='Model Business Impacts Explanation:',
    disabled=False,
    style={'description_width': 'initial'},
    layout=widgets.Layout(height="100%", width="auto")
)
wgt_model_business_explanation
```

Textarea(value='The model provides a moderate estimate of rental prices, with an  $\Box R^2$  score of 0.61, capturing a...

```
[22]: print("Model Business Impacts Explanation:", wgt_model_business_explanation.
```

Model Business Impacts Explanation: The model provides a moderate estimate of rental prices, with an  $R^2$  score of 0.61, capturing a significant portion of rent price variance. However, with an RMSE of 12 on the training set and 22.64 on unseen data, there is room for improvement. The higher RMSE on unseen data suggests that the model may not generalise well, leading to occasional mispricing that could impact revenue and tenant satisfaction. To improve performance, exploring regularised models like Ridge or Lasso could help reduce errors and enhance pricing accuracy.

## 1.7 F. Experiment Outcomes

```
[23]: # @title Experiment Outcomes Explanation

wgt_experiment_outcomes_explanation = widgets.Select(
    options=['Hypothesis Confirmed', 'Hypothesis Partially Confirmed',
    'Hypothesis Rejected'],
    value='Hypothesis Rejected',
    description='Experiment Outcomes:',
    disabled=False,
)

wgt_experiment_outcomes_explanation
```

Select(description='Experiment Outcomes:', index=2, options=('Hypothesis  $_{\square}$   $_{\square}$ Confirmed', 'Hypothesis Partially Con...

```
[24]: # @title Experiments Results Explanation

wgt_experiment_results_explanation = widgets.Textarea(
```

```
value="Hypothesis Partially Confirmed. The model's RMSE of 12 indicates_

that, on average, the predicted rent prices deviate by 12 units from the_
actual values. While this error still indicates some room for improvement,_
it performs better than the baseline model, which suggests that the model is_
capturing meaningful patterns in the data. Reducing the RMSE further would_
result in even more accurate predictions, improving its reliability for_
setting optimal rent prices and minimising potential issues such as lost_
revenue or tenant dissatisfaction.",

placeholder='<student to fill this section>',
description='Experiments Results Explanation:',
disabled=False,
style={'description_width': 'initial'},
layout=widgets.Layout(height="100%", width="auto")
)
wgt_experiment_results_explanation
```

Textarea(value="Hypothesis Partially Confirmed. The model's RMSE of 12 indicates → that, on average, the predict...

```
[25]: print("Experiments Results Explanation:", wgt_experiment_results_explanation.
```

Experiments Results Explanation: Hypothesis Partially Confirmed. The model's RMSE of 12 indicates that, on average, the predicted rent prices deviate by 12 units from the actual values. While this error still indicates some room for improvement, it performs better than the baseline model, which suggests that the model is capturing meaningful patterns in the data. Reducing the RMSE further would result in even more accurate predictions, improving its reliability for setting optimal rent prices and minimising potential issues such as lost revenue or tenant dissatisfaction.