

User Guide for ShinyPET: A Predictive, Exploratory and Text RShiny Application

Contents

User Guide for Shiny PET	2
1. Landing Page	2
2. Exploratory	2
2.1 Observe	2
2.2 Map	3
2.3 Explore and Confirm	6
3. Text	9
4. Predictive	9
4.1 Data splitting	9
4.2 Feature selection	11
4.3 Variables and recipe	13
4.4 Model training	15
4.5 Model evaluation	21

User Guide for Shiny PET

1. Landing Page

Airbnb ShinyPET

- Introduction
- EDA module
- Text module
- Predictive module

Shiny PET: A Predictive, Exploratory and Text Application

About

Shiny PET is a user-friendly application that will enable users to make data-driven decisions without the need to understand programming languages or have extensive statistical knowledge.

We have used Airbnb data as our baseline for this project - data generated is rich in information, which consists of structured, unstructured (textual), and location data.

Application feature

As seen in the figure on the right, this application has 3 modules – exploratory, text and predictive

The Exploratory module allows users to perform exploratory and confirmatory analysis on selected variables to identify interesting patterns.

The Text module allows users to analyse textual data from reviews to generate more quantitative insights.

The predictive module enables users to prepare and build a variety of prediction models.

For more information on this project, please visit our [website](#)

Overview of ShinyPET

Exploratory	Text	Predictive
Observe	Word Frequency	Data Splitting
Map	Sentiment	Feature selection
Explore & Confirm	Topic modeling	Data Transformation
	Network Analysis	Model training
		Model Evaluation

Figure 1: Landing page of Shiny PET

The landing page of the application provides a brief background for this application, its features and overview of its navigation.

2. Exploratory

2.1 Observe

This tab allows user to quickly understand the data to be analysed.

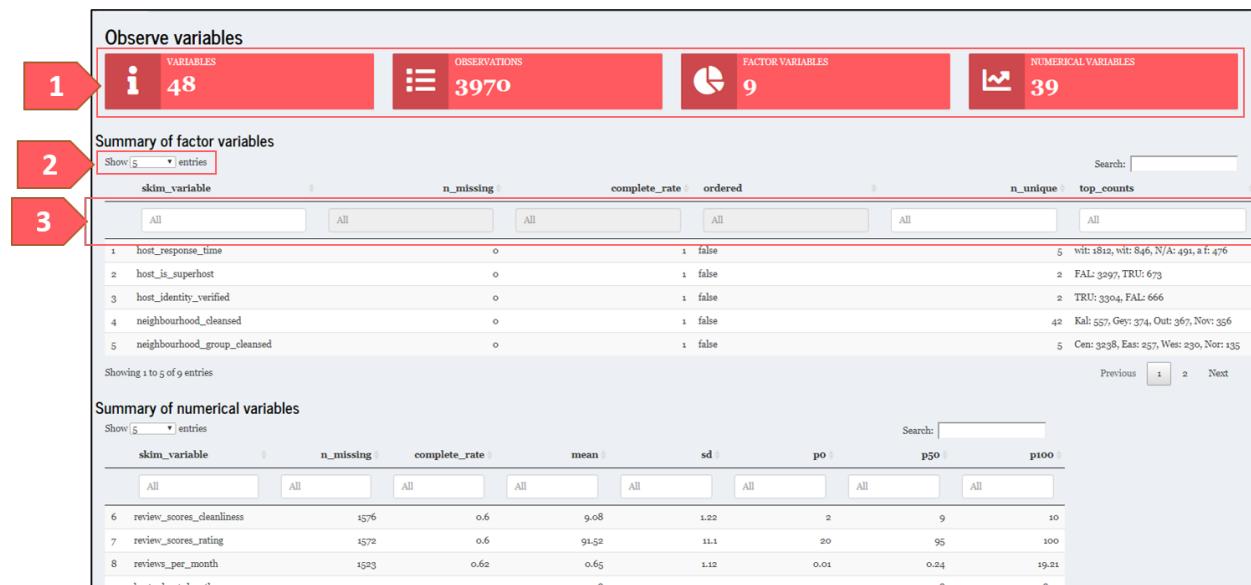


Figure 2: Observe tab of Explore module

[1] Shows the summary of data.

[2] Change number of observations shown. To minimize having to constantly select “Next” to view hidden variables, users can show the maximum entries (i.e 100) to be displayed.

[3] Search data if necessary.

2.2 Map

This tab allows user to explore the geographic patterns of Airbnb listings through 2 thematic maps - point symbol and choropleth.

2.2.1 Point Symbol map Each point on the map is a listing. This allows user to see how distributed Airbnbs are throughout Singapore.

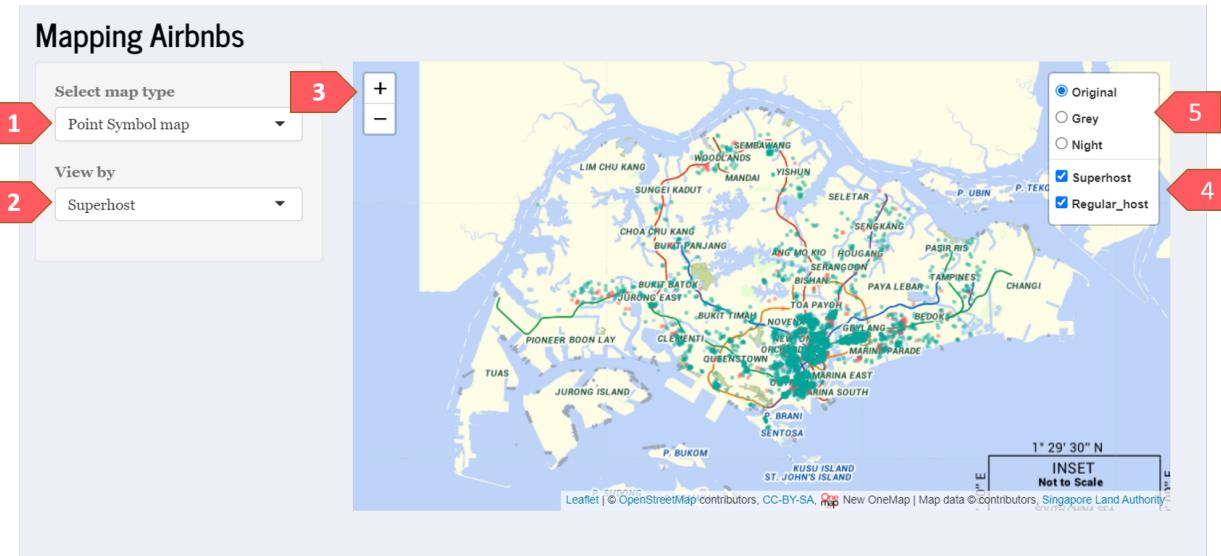


Figure 3: Point Symbol Map

- [1] Select map type. Map will auto update upon selection.
- [2] Select between superhost and room type. This shows the listings by selected variables.
- [3] Zoom in and out the map.

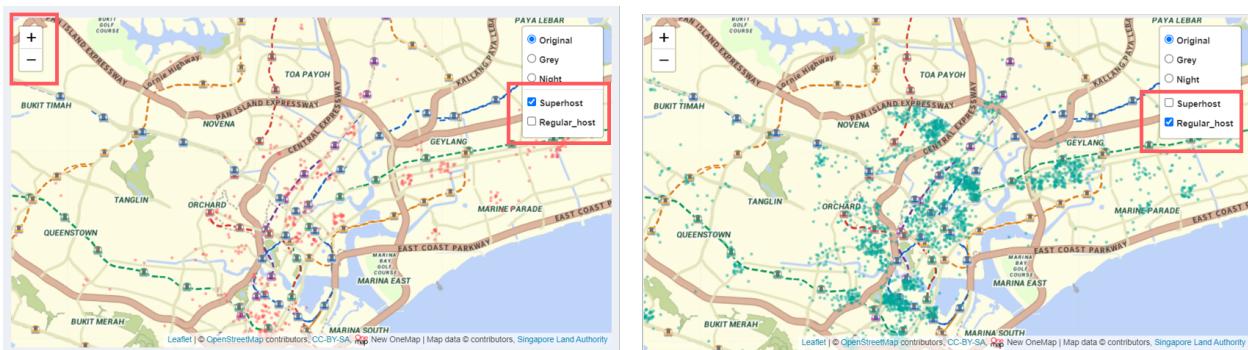


Figure 4: View selected listings

- [4] User can check and uncheck to view selected listings.

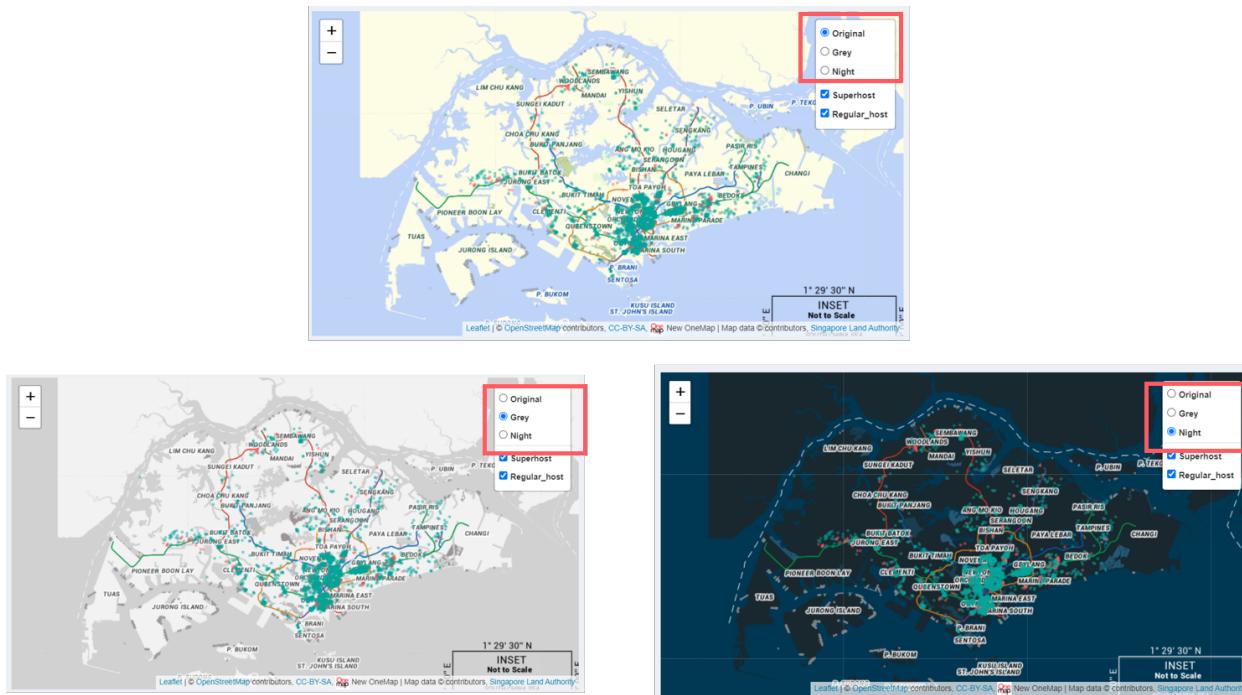


Figure 5: 3 types of background colour

[5] User can select between different background.

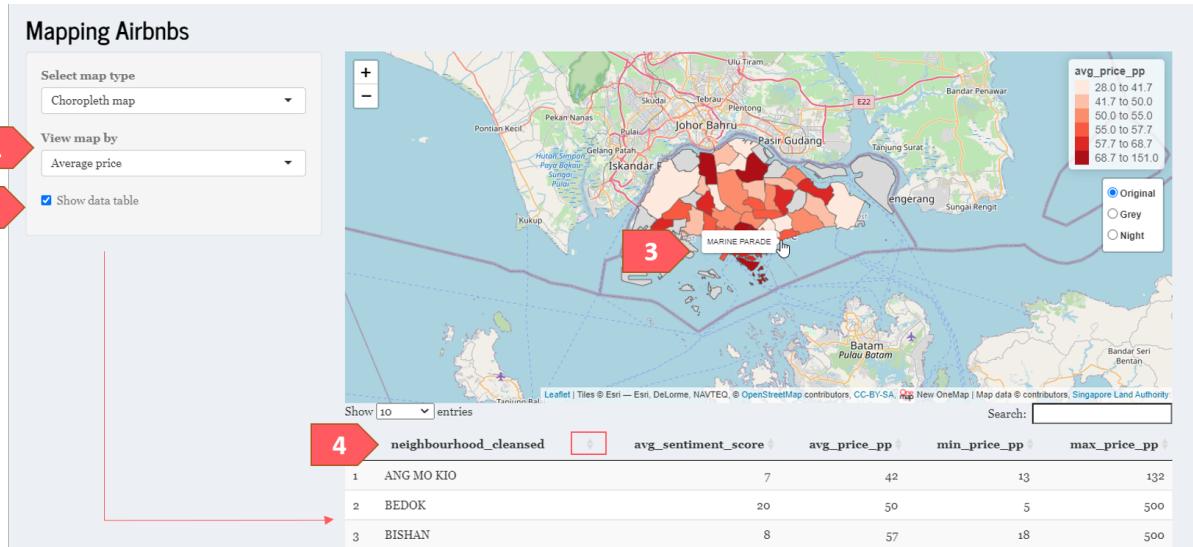


Figure 6: Choropleth map

2.2.1 Choropleth map [1] Select between average price and average sentiment score. The average sentiment score is calculated based on the positive

and negative of guests' reviews. For more details on sentiment of reviews, please fer to the Text module > Sentiment Analysis tab.

[2] User can decide whether to view the data by checking / unchecking the 'Show data table' box.

[3] Hover over the area to see neighbourhood names.

[4] User can sort table according to needs.

2.3 Explore and Confirm

This tab allows users to perform exploratory and confirmatory analysis.

2.3.1 Types of chart available User can explore the dataset using different charts.

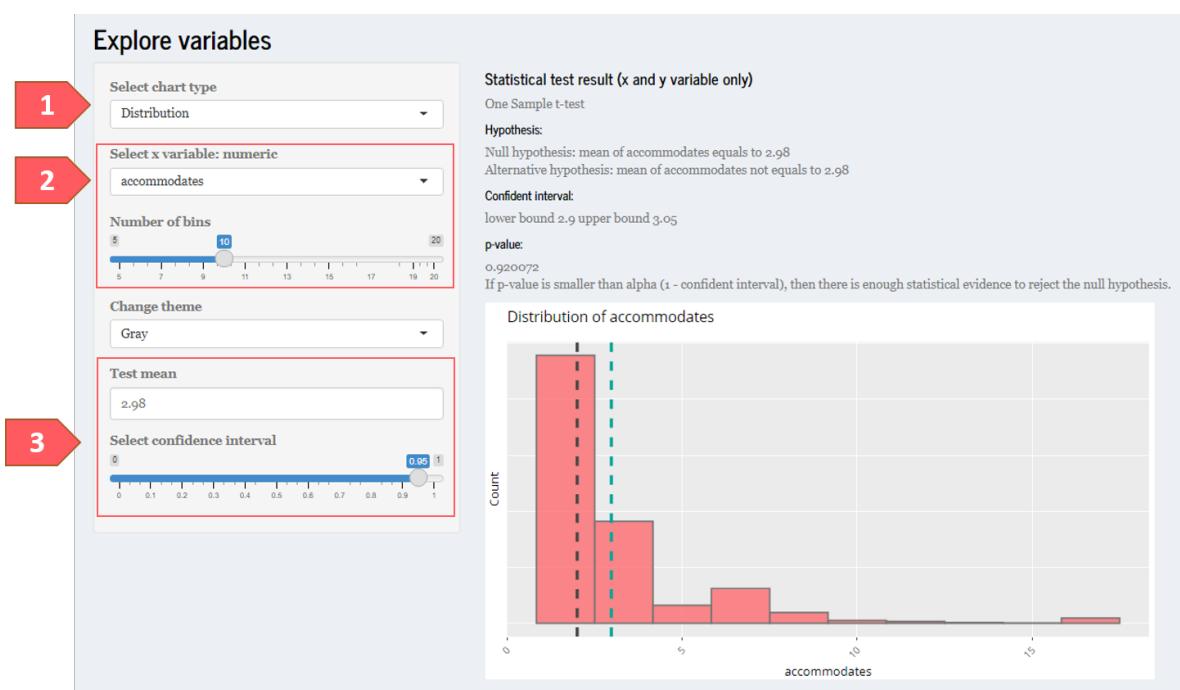


Figure 7: Explore and Confirm tab of Explore module

[1] There are 4 types of chart:

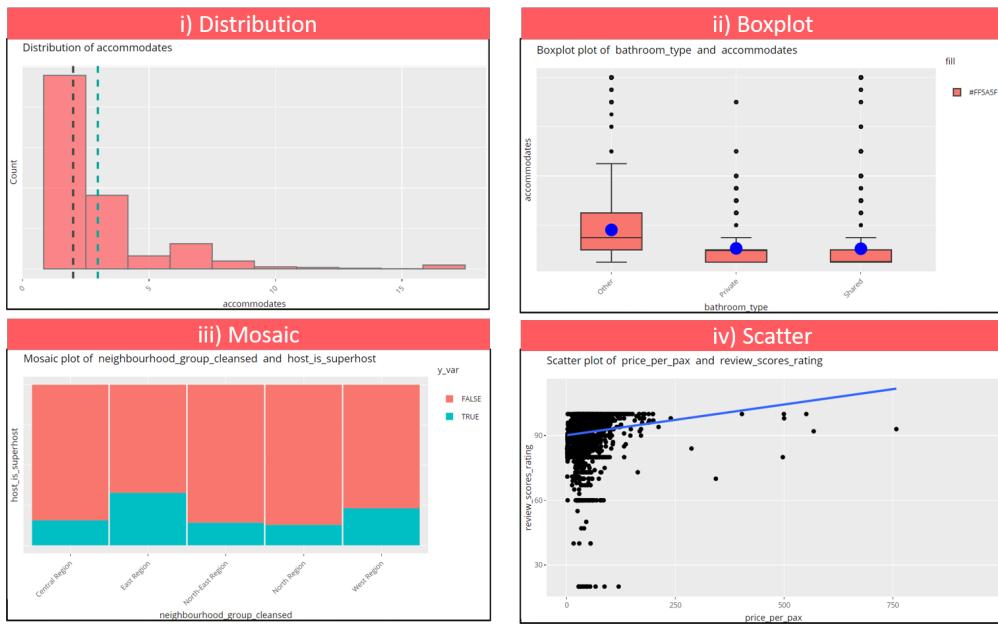


Figure 8: Four types of charts available

i) Distribution - to analyse a single variable (univariate analysis).

- teal line refers to mean
- black line refers to median

ii) Boxplot - to analyse 1 factor and 1 numeric variables

- Blue circle refers to mean

iii) Mosaic - to analyse two factor variables

iv) Scatter - to analyse two numeric variables

[2] Selection available will change according to chart type.

[3] Statistical test options available will change according to chart type.

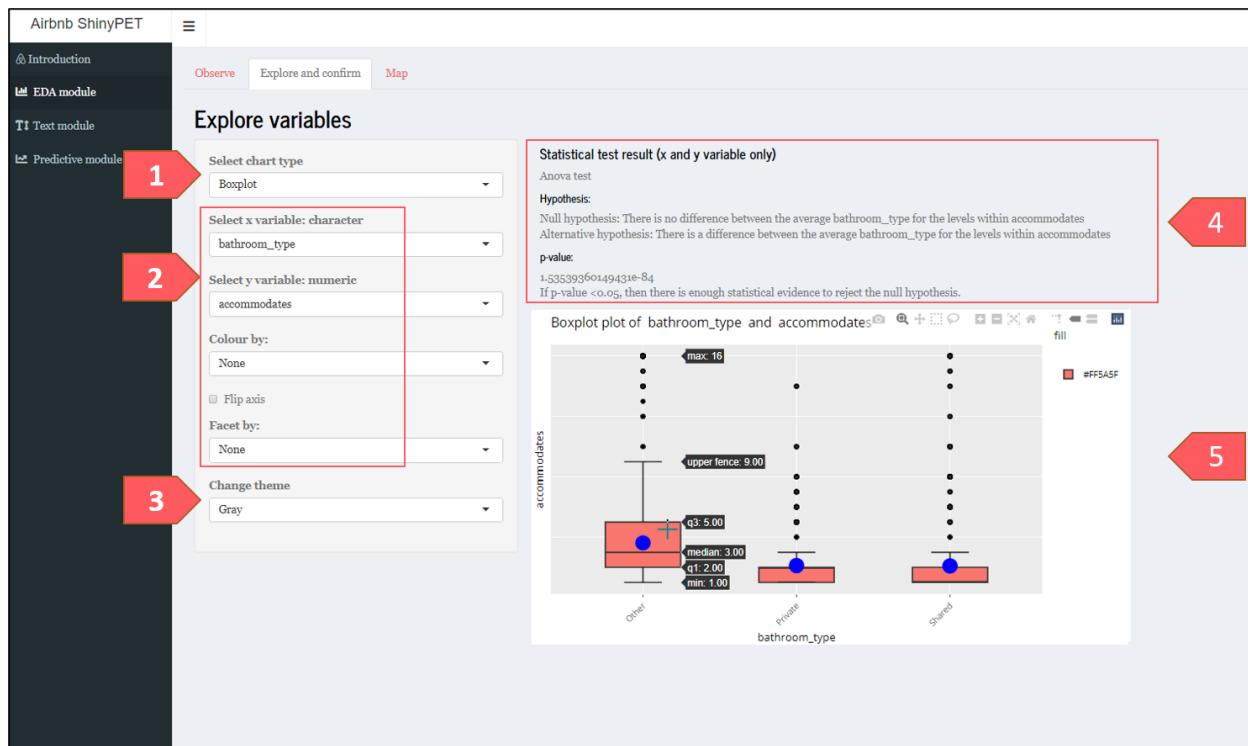


Figure 9: Explore and Confirm tab of Explore module

2.3.2 Performing exploratory and confirmatory analysis [1] Select chart type.

[2] Select the variables that you are interested in analysing.

[3] This changes the background of the graph.

[4] The type of statistical tests are automated based on user's x and y variables. Simply select the variables you wish to analyse.

If p-value is less than the alpha (1 - confident interval), you reject the null hypothesis and accept the alternative hypothesis.

Note: Statistical test is only applicable to selected x and y variable, and does not take colour and facet variables into consideration.

[5] The chart is interactive and users hover to select a single object in a plot, highlight selected records by clicking and unclicking legend, define a region and download the chart.

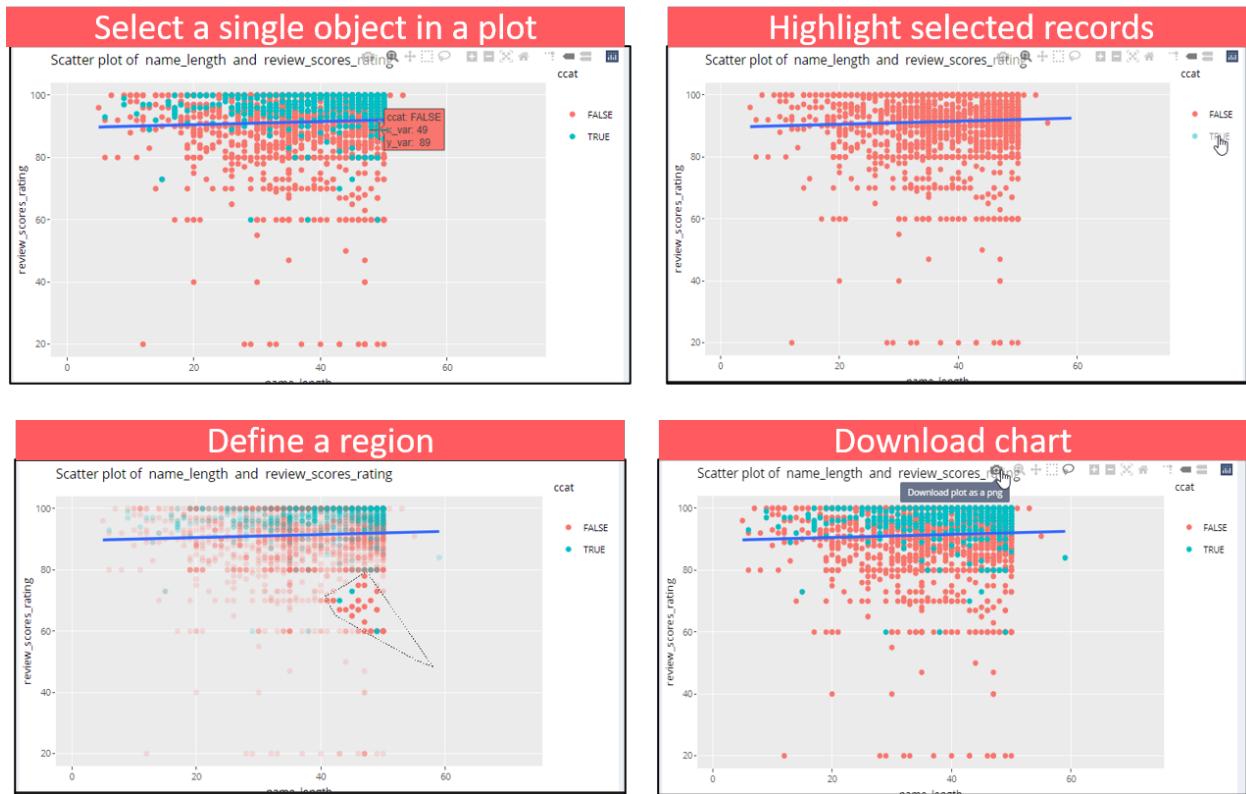


Figure 10: Interactive chart

3. Text

4. Predictive

4.1 Data splitting

The module starts with the tab which allows user to:

- [1] Select the response variable for prediction
- [2] Choose the proportion of training and validation (test) set
- [3] Proceed with data splitting

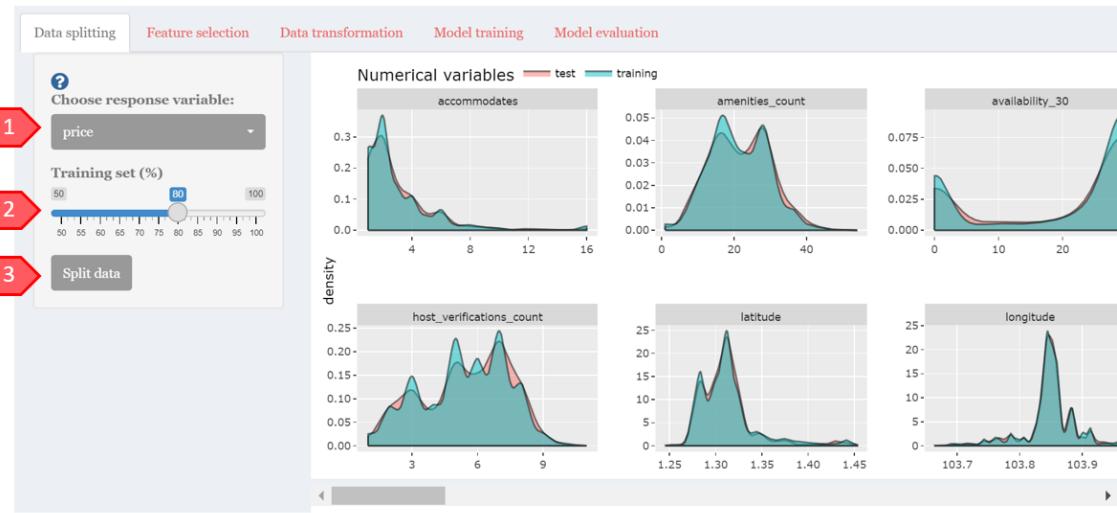


Figure 11: Data split tab of Predictive module

This will then create density plot for numerical variables [4] and bar chart for categorical variables [5] with division between training and test set.



Figure 12: Train-test split distribution plot

4.2 Feature selection

User will be presented with 2 options for feature selection process: correlation matrix and feature selection using Random Forest and Boruta method. In the correlation matrix section, user can:

- [1] Select variables of choice
- [2] Select correlation method
- [3] Select p-value significance criteria
- [4] Produce correlation matrix
- [5] Hover over matrix to view details on correlation value and p-value

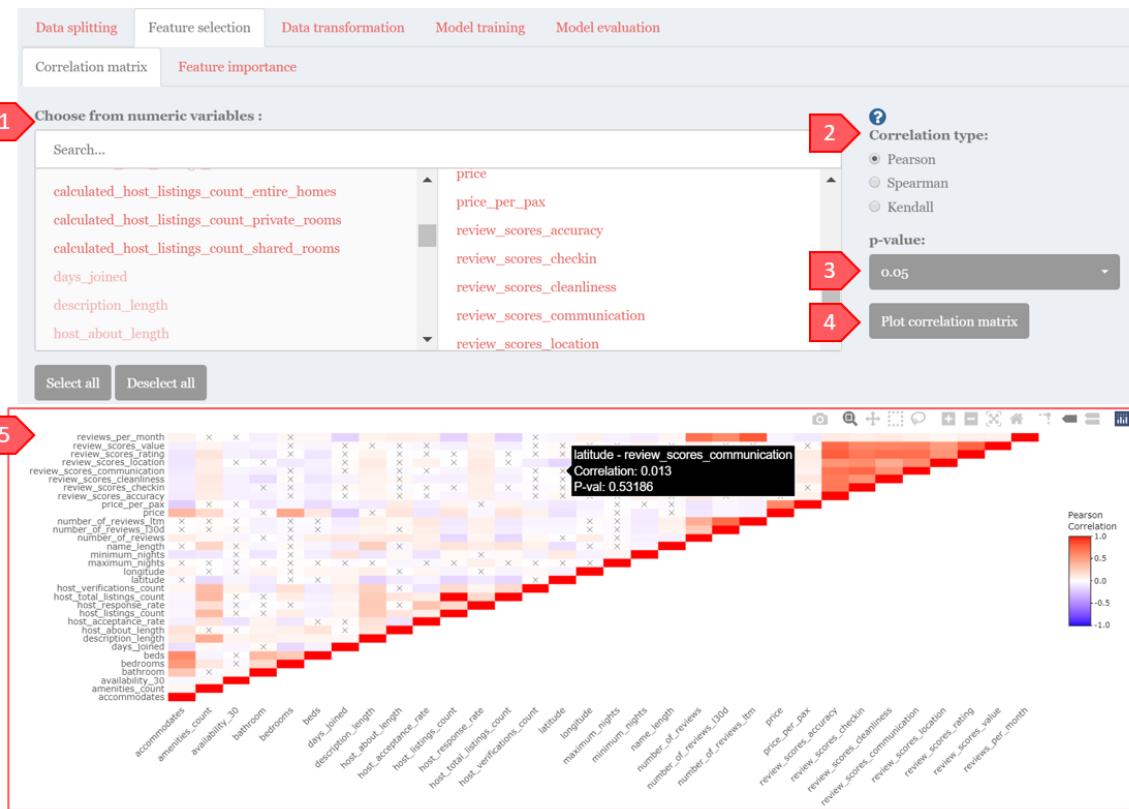


Figure 13: Correlation matrix tab of feature selection

In the next section, user may decide to run the feature selection process by Random Forest and Boruta method using the training set created from the previous data splitting section. The plot for both methods will be displayed side by side for comparison.

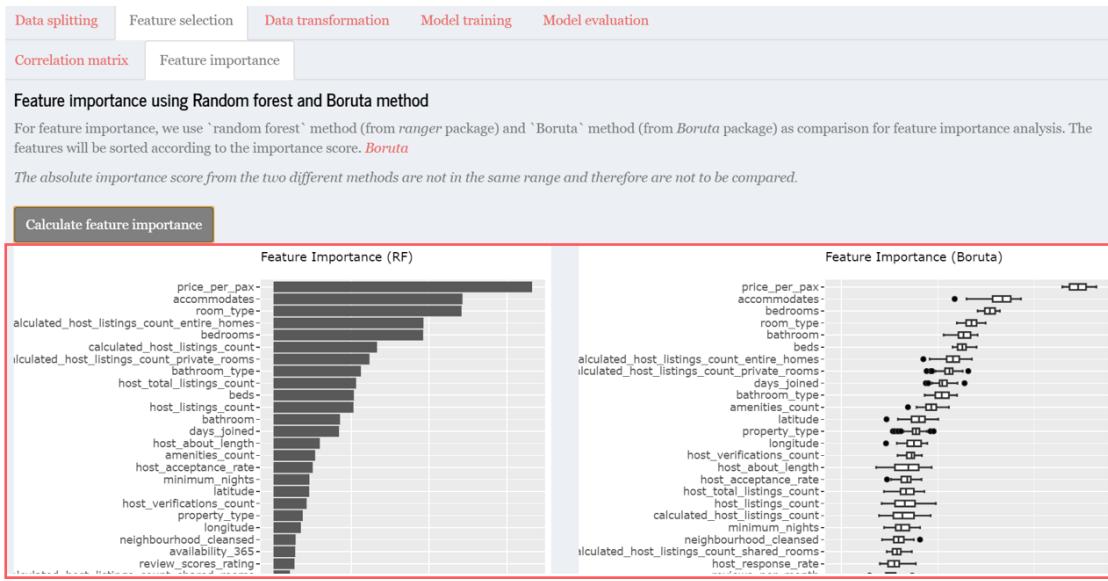


Figure 14: Feature selection tab comparing 2 different methods

4.3 Variables and recipe

Based on the result in the feature selection section, this tab allow user to filter the variables to be included in the predictive model. This can be done through the multi-input selection form [1]. This module comes with predefined data transformation steps which can be displayed by clicking on the “Prepare recipe” button [2]. This will provide user with the recipes that will be executed in data transformation prior to model training [3]. To finalise variable selections and apply the transformation steps, user can click on the “Transform variables” button [4] which will navigate the page to the next section.

1 Variables and Recipe 2 Prepare recipe

3

4 Transform variables

Figure 15: Variable selection and data transformation tab

In this section, user will be able to check the train-test split distribution once again after variable selection.

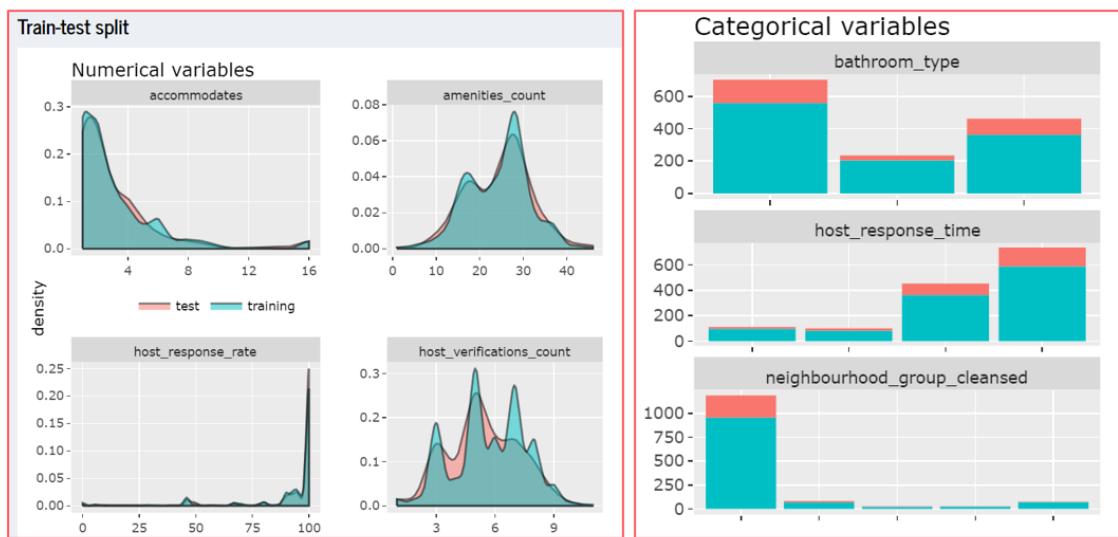


Figure 16: Train-test split distribution after variable selection

The transformed numerical variables are also displayed for comparison, with option to select variable of interest [5].

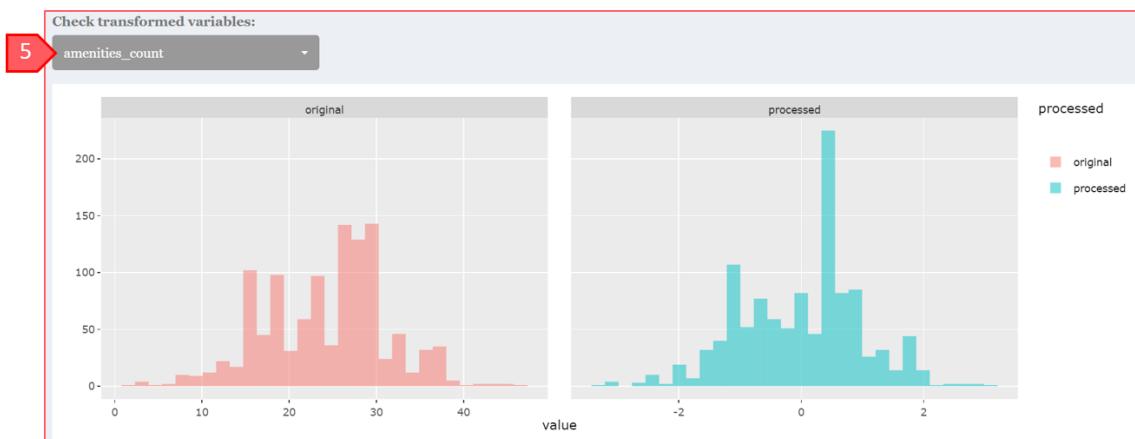


Figure 17: Transformed variables before and after comparison

4.4 Model training

To train the model, user is provided with a selection of model algorithm e.g. linear model and tree-based model. Each of the model algorithm has its own tabs which mainly comprises the following set of sections: information of the model, training result, and validation result.

4.4.1 Model information This section provides background of the model that is going to be trained. External source is given in hyperlink for user to learn more information about the model [1].

A. Training linear model User can proceed with model training by clicking the “Train model” button [2].

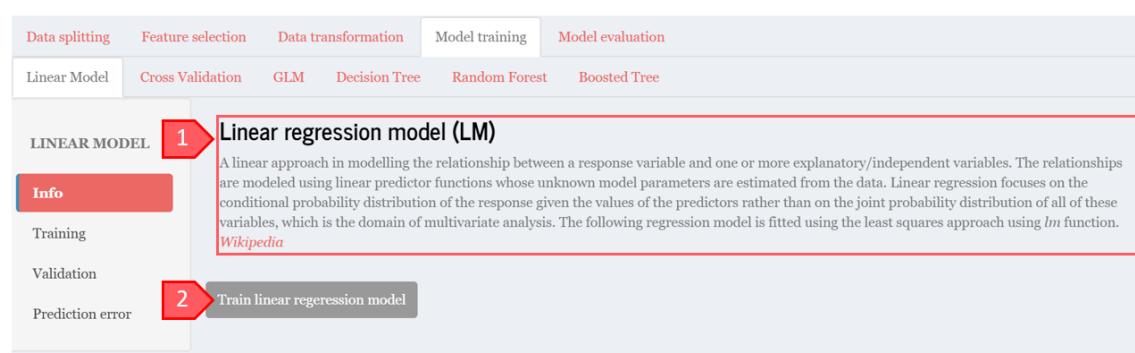


Figure 18: Model information page and training

B. Training other models Other models (Generalised Linear Model, Decision tree, Random Forest, and Boosted Tree) require cross validation (CV) training sets. Prior to model training, user will be required to:

- [1] Navigate to the “Cross Validation” page
- [2] Select K-fold cross validation parameter
- [3] Prepare cross validation training set

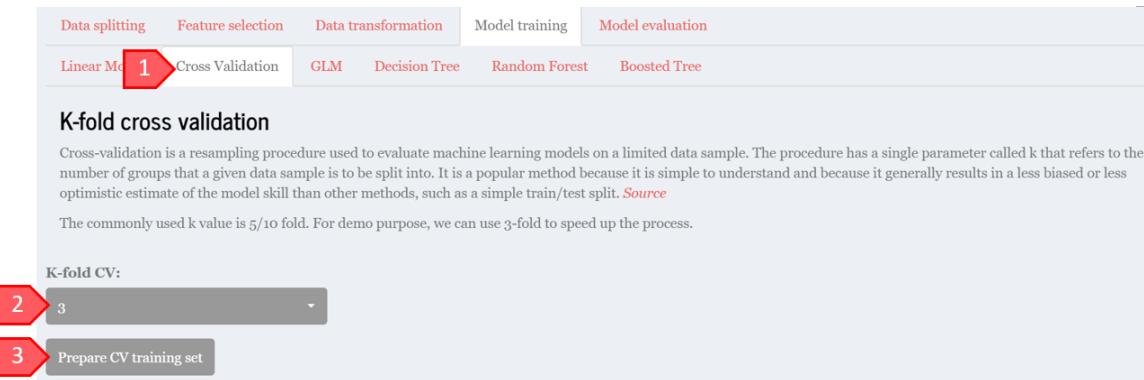


Figure 19: K-fold cross validation preparation

Once the CV training set is prepared, user can proceed to train the rest of the models. Note that training for these models will take some time and user will be prompted by a message which will disappear once model training is completed.

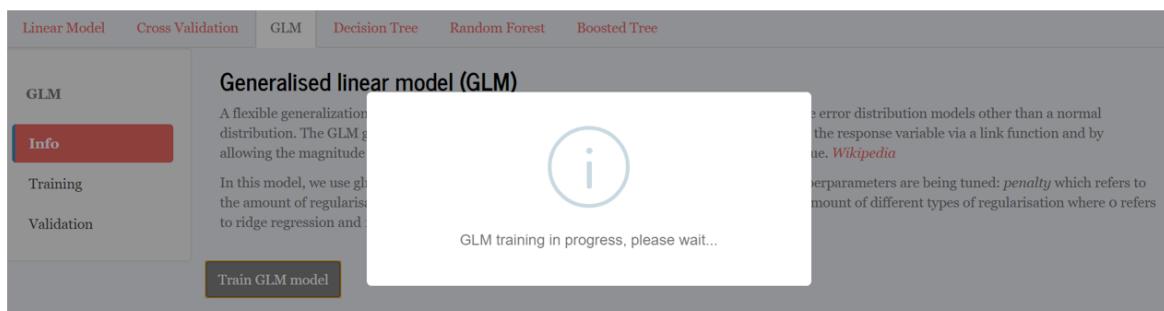


Figure 20: Training in progress loading page

4.4.2 Training result Once model training is completed, user will be directed to the next section for the training result.

A. Linear model For linear model, the model fit result is given in a table form [1]. Interactive exploration on coefficient estimate result is possible by

selecting p-value criteria [2], sorting method [3], and tooltip information [4]. User is to proceed with model validation by clicking on the “Validate model” button [5].



Figure 21: Training result page of linear model

B. Other models The rest of the models’ training involves hyper-parameter tuning where the summary will be displayed once training is completed.

[1] The hyper-parameters are plotted and grouped according to 4 types of metric (RMSE, MAE, MAPE, Rsquared)

[2] Next, user will be able to view the best model according to the metric of choice through the dropdown menu selection.

[3] The data table below will respond to the metric selection and display the best model on top

[4] Click the “Choose best model” button to select the best model based on selected metric.

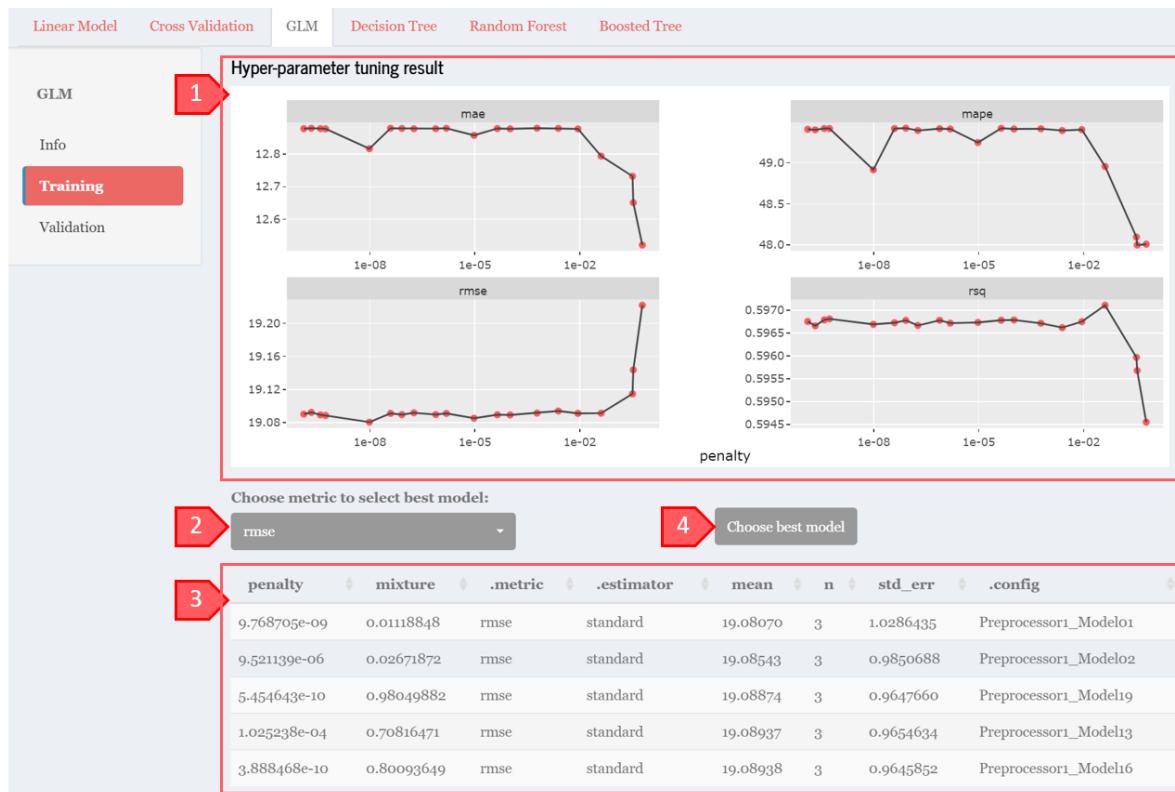


Figure 22: Training result page for GLM

Once best model is selected, different information will be displayed depending on the model algorithm. For GLM, the coefficient estimate will be plotted and sorted according to the estimate value.

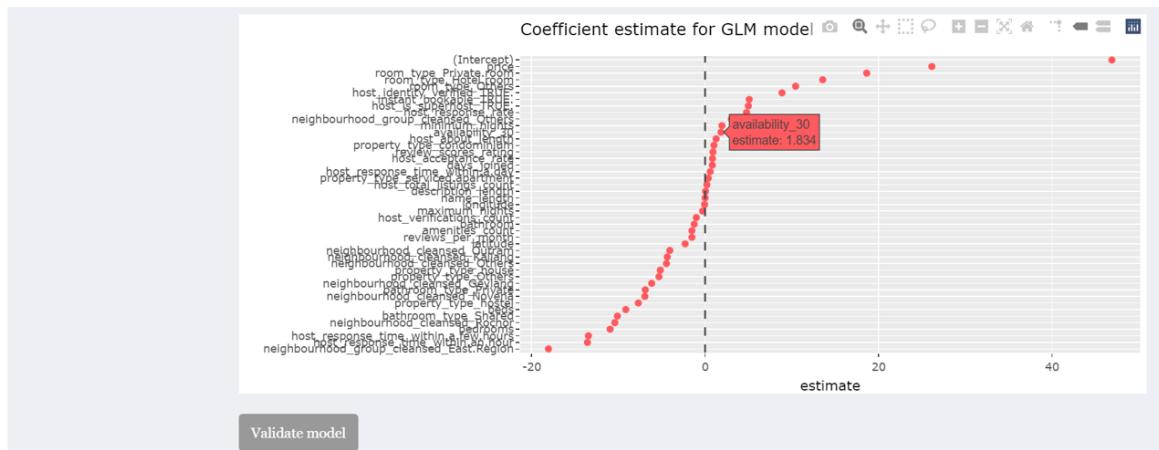


Figure 23: Coefficient estimate plot

Tree based model will have variable importance score displayed, with additional tree visualisation for decision tree model.



Figure 24: Variable importance and tree visualisation

4.4.3 Validation result After validation process, the predicted value will be plotted against the actual value in an Rsquare plot [1]. The calculated metric performance is also provided below the plot [2].

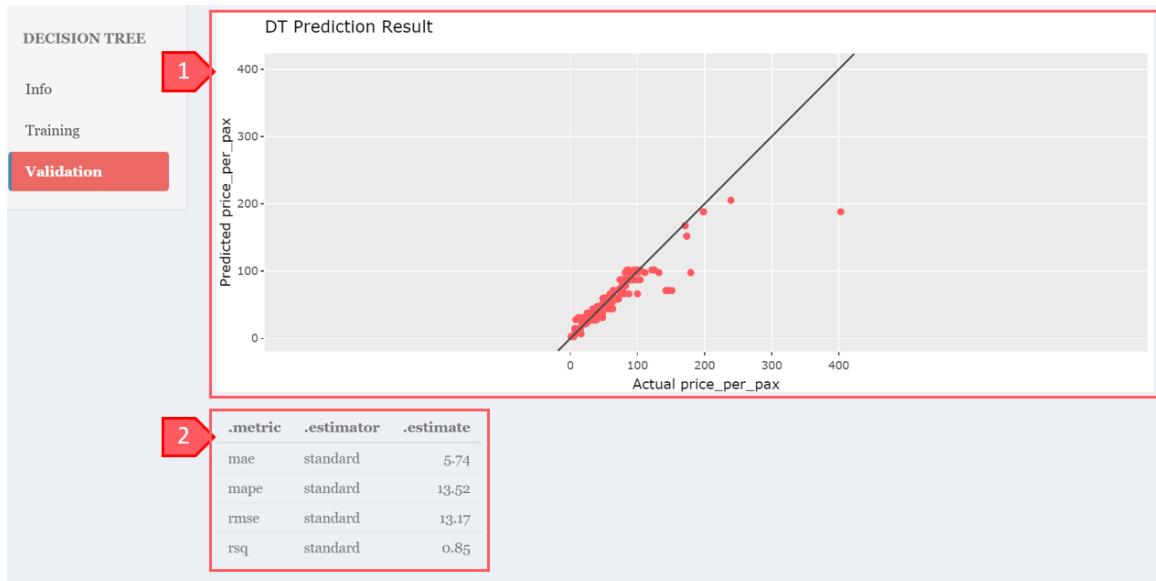


Figure 25: Model validation result page

4.4.4 Prediction error For linear model, additional section is available to explore cases with high prediction error.

- [1] Select how many prediction points with highest deviation from actual value
- [2] Select how many top predictors to be displayed
- [3] Select p-value threshold for deciding significant predictors
- [4] Plot of training data distribution as histogram, overlapped with points where prediction error is high
- [5] User can single out specific point by double clicking the “id” number in the legend
- [6] Details of point with high prediction error and their predictors value are displayed in a table

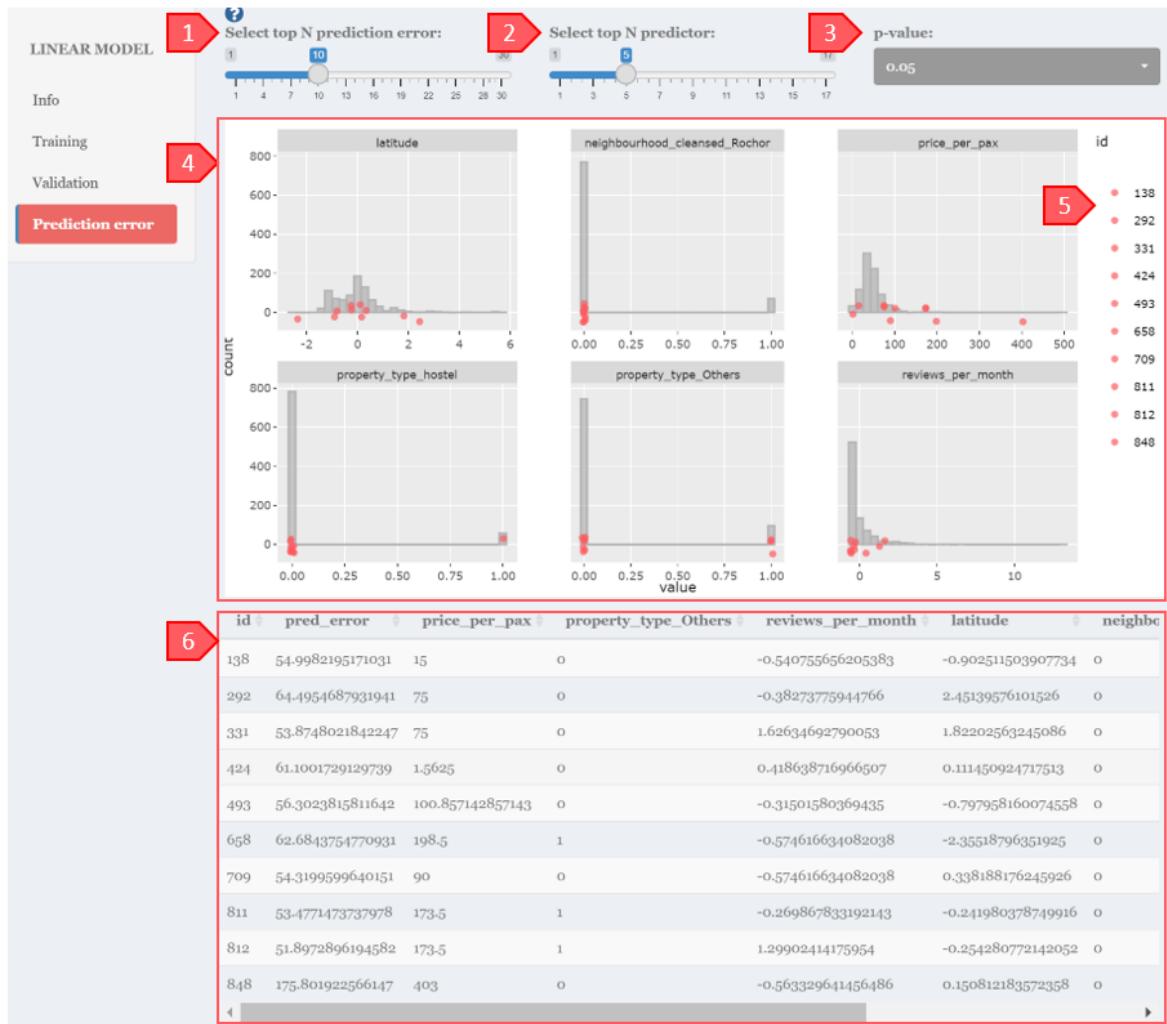


Figure 26: Prediction error page for further exploration

4.5 Model evaluation

To perform model evaluation, ensure that at least 1 of the model algorithms has been trained and validated through the previous sections. This section comprises 3 pages:

4.5.1 Information page To start the evaluation, click on the “Collect model performance” button to gather all the best model that has been trained and validated previously.

Figure 27: Model evaluation page

- 4.5.2 Performance** Upon clicking the button, the next page will show:
- [1] Comparison of models along with their performance metrics
 - [2] Through the dropdown menu, user can select the best model that will be used for prediction
 - [3] Finalise model selection by clicking the button



Figure 28: Model performance comparison page

- 4.5.3 Predict** After selecting the final model, user will be directed to the next page for prediction with new input variables.

- [1] All variables selected for model training are displayed for user input
- [2] Once input variables are defined, click on the button to calculate response variable using the selected model
- [3] Predicted value is displayed

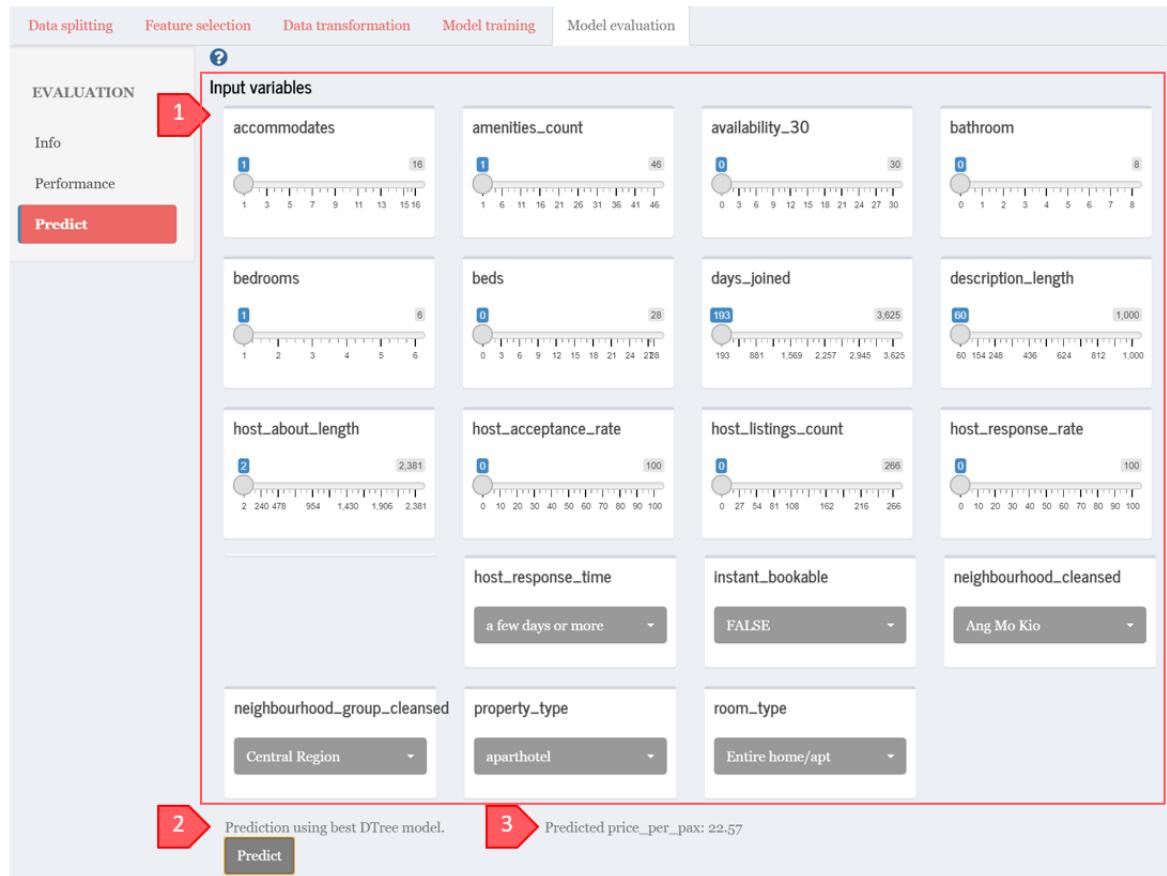


Figure 29: Prediction page with input variables and predicted value