# An Analysis of Marketing Campaign Dataset

*STDS AT3 Project Report*

Thanh Nhan (Nicholas) Le – 11919925
Statistical Thinking for Data Science
TD School
University of Technology Sydney

UTS CRICOS 00099F

# Contents

### Abstract

This project aims to identify customers most likely to subscribe to a new telecommunication plan, and to give expert advice to management based on analysis insights. To this end, statistical models (Logistic Regression (LR), k-Nearest Neighbours (kNN), Decision Tree (DT), and Random Forest (RF)) were trained and cross-validated to achieve the best f1-score, whilst maintaining a high recall. A realistic campaign simulation was run to decide on the best model based on profitability achieved in different market conditions. LR and RF were found to be comparable in cross-validated performance, and although RF achieved the lowest f1-score because of low precision, it had the highest recall. Simulation results pointed to RF as the winning model in profitability. RF achieved a recall of 95% on the test set, demonstrating a superior ability to identify buying customers. Insights for management include monitoring call qualities and macroeconomic factors such as employee variation rates, Euribor rates and consumer price index, which may affect buying potentials.

# 1 Problem formulation

This project extends on the last exploratory data analysis (EDA) of a marketing campaign dataset. It has two clear sets of goals:

1. **Technical goal**: Identify customers that are most likely to subscribe to the new telecommunication plan. It achieves this goal by training a mix of parametric, non-parametric and ensemble statistical models to maximise the f1-score, while maintaining the highest recall. Recall represents the ability of the model to identify customers with an intention to buy, whilst the f1-score maintains the precision-recall balance in model training.
2. **Business goal**: Devise strategies for management that are directly applicable in a campaign setting. This is done through interpreting model insights, such as parametric coefficients inference and feature importance interpretation.

# 2 Data preprocessing

## 2.1 Imputation: `unknown` values

There were 6 categorical columns with `unknown` values, which were imputed using v. Buuren & Groothuis-Oudshoorn (2011)'s multivariate Bayesian imputation method using chained equations. The distributions of these columns remained almost the same post-imputation.

## 2.2 Binarisation

Categorical variables with 3 or more unique categories are binarised, yielding 0/1 columns for each category. Using this encoding method preserves information, making model interpretations easier.

## 2.3 Data splitting

The dataset is heavily imbalanced, with almost 90% of observations classified as non-subscribers. A stratified train-test split was performed to preserve class ratios, and the test size was 20% according to industry practice.

## 2.4 Data scaling

Scaling of the whole dataset, including 0/1 variables, is needed for the Lasso-penalised LR, as advised by Tibshirani (1997). The standard scaler was used by subtracting each data point $x_{ij}$ by column $j$'s mean $\bar{\mathbf{x}}_{j,\text{train}}$, then divide by the standard deviation of $\mathbf{x}_{j,\text{train}}$:

$$\dot{x}_{ij} = \frac{x_{ij} - \bar{\mathbf{x}}_{j,\text{train}}}{\text{std}(\mathbf{x}_{j,\text{train}})} \tag{1}$$

# 3 Statistical modelling

Each of the following models will be validated for performance using a 5-fold cross-validation strategy on the training set. This method reduces overfitting, while allowing the generalisation power of the models to be evaluated. The best model will undergo a final round of testing on the test set.

The scoring metric is the f1-score, to handle the imbalanced dataset:

$$\text{f1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

## 3.1 Model choice

### 3.1.1 Parametric

Logistic Regression (LR) was chosen because of its straightforward interpretation of coefficients. In the last project, EDA suggested that this would be a good starting point for modelling, because of some linear relationship found between the predictors and the response.

Lasso will be used to prevent overfitting, since binarisation significantly inflated the number of covariates in the design matrix. The goal is to minimise the Lasso-penalised negative log-likelihood via maximum likelihood estimation:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \big[ - y_i \ln(\hat{p}(\mathbf{x}_i)) - (1 - y_i) \ln(1 - \hat{p}(\mathbf{x}_i)) \big] + \lambda \|\boldsymbol{\beta}\|_1, \tag{3}$$

where $\hat{p}(\mathbf{x}_i) = 1 \big/ (1 + \exp -(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0))$ is the estimated probability that the observation $\mathbf{x}_i$ belongs to the positive class.

### 3.1.2 Non-parametric

k-nearest neighbours (kNN) was chosen because of its simplicity and ease of interpretation. Its goal is to predict a class based on neighbour data points, by finding the class $c$ that maximises the number of neighbour points $\mathbf{x}_j$ in a set $S_k$ of $k$ neighbours:

$$\hat{y}_i = \underset{c}{\operatorname{argmax}} \sum_{j:\, \mathbf{x}_j \in S_k} \mathbb{1}\{\mathbf{x}_j \text{ in class } c\} \tag{4}$$

Decision tree (DT) was also considered in addition because of its computational speed, while also providing a base for training ensemble algorithms.

#### 3.1.3 Ensemble

Random Forest (RF) was built as an ensemble of multiple DT's to enhance the overall predictive power, whilst reducing the effect of overfitting by the individual trees. It predicts a class based on the majority vote of $n$ DT's, where $n$ is to be determined via cross-validation.

### 3.2 Summary of model configurations

The following optimal hyperparameter sets for each model were obtained via cross-validated randomised search. Unmentioned hyperparameters are assumed to take default values as per the scikit-learn API (Pedregosa et al. (2011)).

| Model | Hyperparameters |
|---|---|
| LR | • Class weight: "balanced"<br>• Max iteration: 500<br>• Penalty: "l1"<br>• Solver: "saga" |
| kNN | • Number of neighbours: 50<br>• Weights: "distance" |
| DT | • Class weight: "balanced"<br>• Max. depth: 22<br>• Max. features: 7<br>• Min. samples leaf: 660<br>• Min. samples split: 970 |
| RF | • Class weight: "balanced"<br>• Max. depth: 22<br>• Max. features: 7<br>• Min. samples leaf: 660<br>• Min. samples split: 970<br>• Number of estimators: 110 |

**Table 1:** Summary of model hyperparameters.

# 4 Analysis of results

Table 2 provides the 5-fold cross-validation results for each model in the training set.

| Model | Cross-validation f1-score | True positive rate | False positive rate |
|---|---|---|---|
| LR | 0.5948 | 0.8889 | 0.1397 |
| kNN | 0.5414 | 0.4787 | 0.0367 |
| DT | 0.5178 | 0.8480 | 0.1809 |
| RF | 0.4956 | 0.9094 | 0.2234 |

**Table 2:** Summary of cross-validation results.

Based on these, LR and RF are comparable in performance: LR is well-rounded in predictive power; but RF is the highest in terms of true positive rate, meaning it is the best model in detecting buying customers.

A campaign simulation and cost sensitivity analysis was run to decide between these models, with assumptions made about costs and revenue per customer. The winning model is one that makes the most profits. Details of campaign assumptions and sensitivity analysis are available in Appendices A and B.

Figure 1 provides the sensitivity analysis results, from which it is evident that RF performs better than LR in all market conditions. Axes represent critical cost factors, and the colour map is the profit difference between RF and LR.

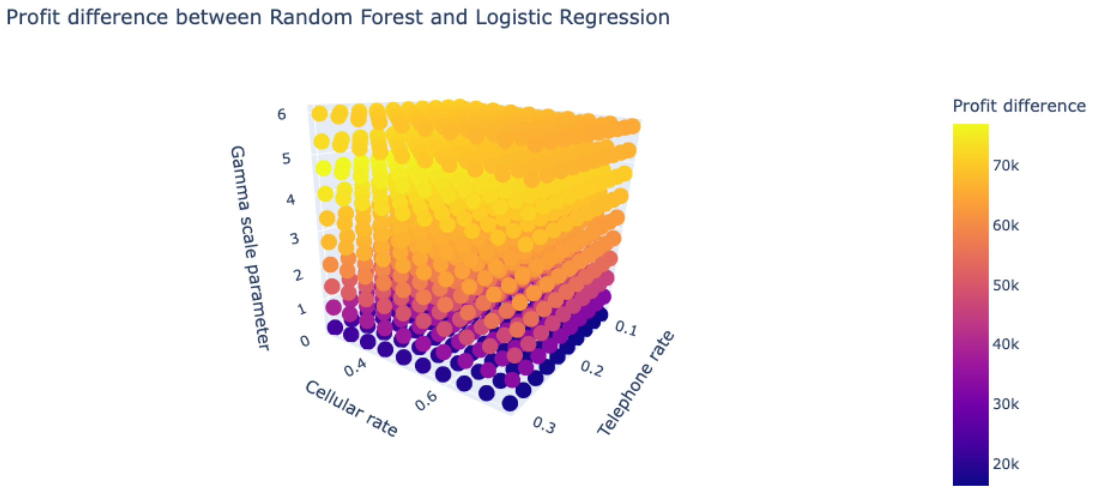

**Figure 1:** Sensitivity analysis of cost factors on the models' profit difference via a colour plot. Fully interactive plot available in Appendix C.

RF was chosen as the final model for the business, because the profit difference is consistently positive. Figure 2 shows its generalisation power on the test set. Its positive predictive power was satisfactory.

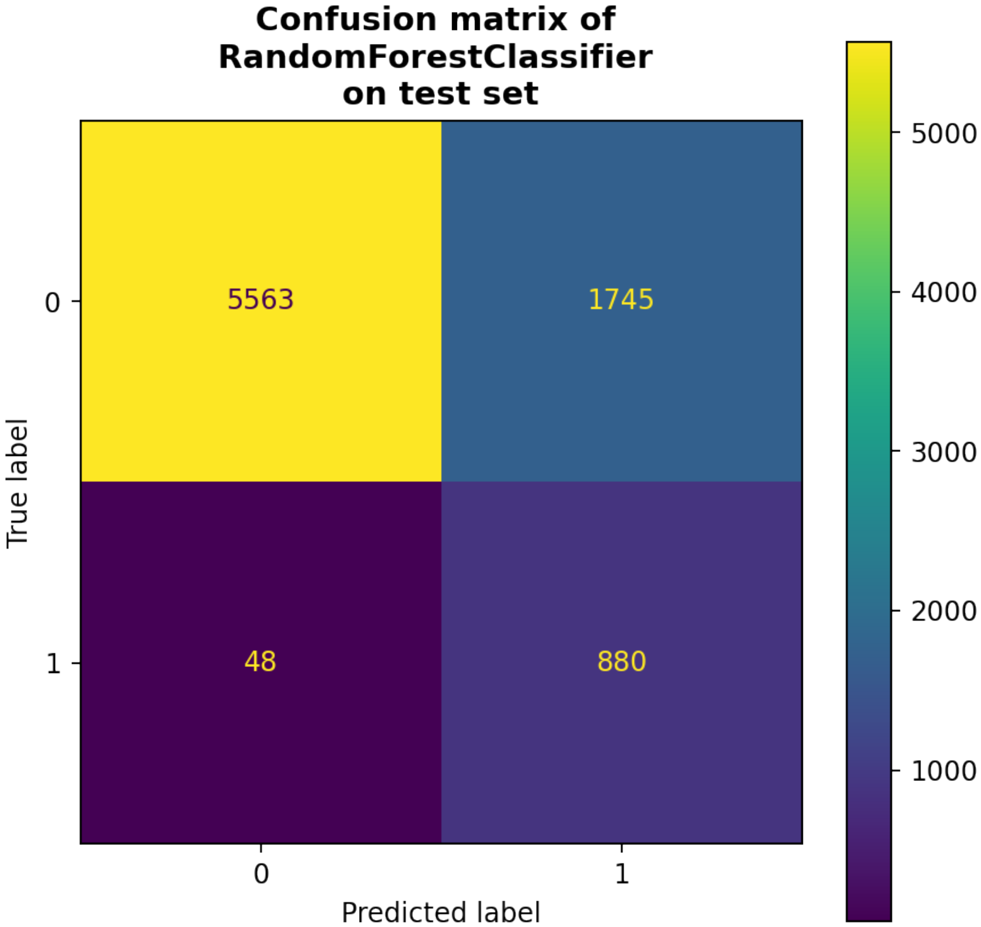

**Figure 2:** Confusion matrix of Random Forest on the test set. It achieves a 95% recall.

# 5 Insights

1. Figure 3 shows that the groups most receptive to this campaign are young students in their 20s, and retired people above 60. Further, Figure 4 points to groups that are more likely to subscribe:

   Blue collar workers

   Married people

   Those in higher education

   Those in generally good credit standing (no defaults, no personal loans)

2. Management to pay close attention to call quality. This is reflected in duration of calls, and is the most effective indicator of customer subscription. Figure 5 reflects its

positive direction of movement with the response variable in the LR, while Figure 6 shows its largest contribution to the mean impurity reduction in the RF. Longer calls usually signify more quality conversations with customers.

3. Management to proactively respond to macroeconomic changes, which may affect subscription. Figure 5 shows that employment variation rates most negatively affects subscription levels.

4. Macroeconomic factors like the 3-month Euribor rate and consumer price index also have major contributions to subcription levels, as shown in Figure 6. These call for regular reviews of product pricing and services offered, in order for business to stay competitive and maintain/increase customer retention rates.

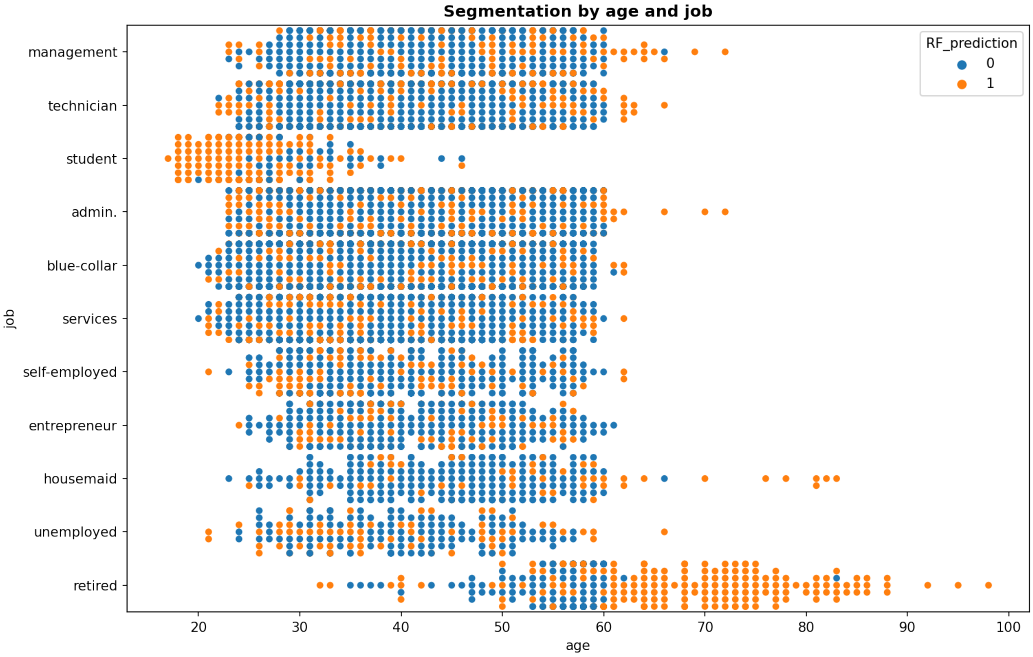

**Figure 3:** Customer segmentation by age and job.

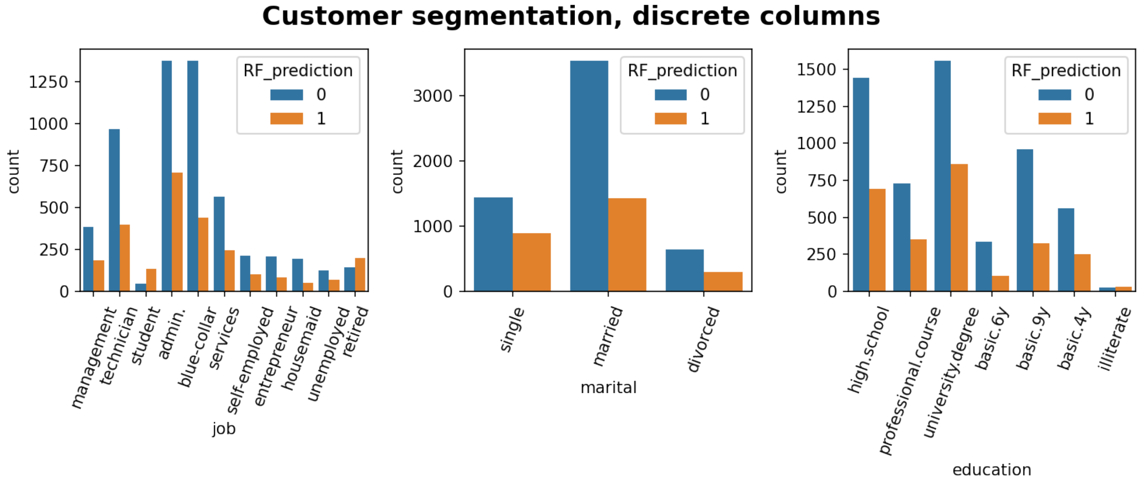

**Figure 4:** Customer segmentation by categorical variables.

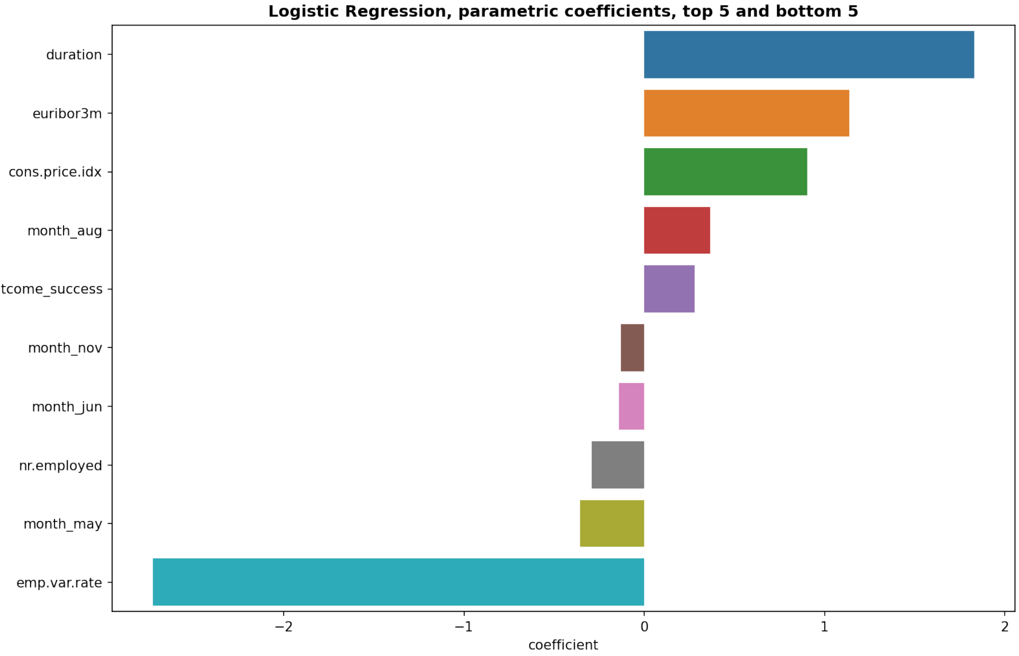

**Figure 5:** Parametric (standardised) coefficients of Logistic Regression, top 5 and bottom 5 in magnitude terms.

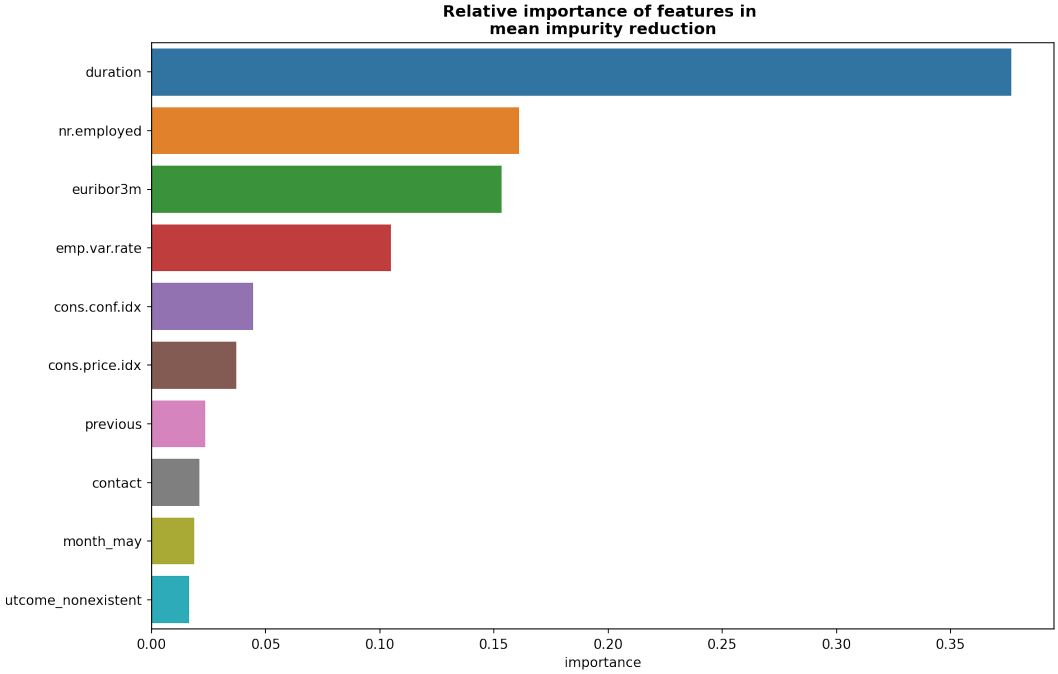

**Figure 6:** Feature importances, inferred by Random Forest.

# References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, *16*(4), 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

v. Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. doi: 10.18637/jss.v045.i03

# Appendices

## A Marketing campaign simulation details

Assume the following yearly costs for an inbound telemarketing campaign, which is a realistic case for this dataset. The landline services are offered by **Cadiz3**, which is a market competitive company:

- Fixed rate: \$600/12 months minimum total cost
- Call rates: 10c/min telephone, 25c/min mobile. All rates are ceil rates (calls less than a minute are charged for a full minute)
- Since call durations are available, all customers are assumed to have picked up the call
- There are no other costs, for simplicity.

Also, assume the telecom plan to sell is an **nbn25** plan from Optus, which is a base telecom plan:

- If customer onboard is successful: the customer may stay for a variable number of months, $x$, which is assumed to be a continuous random variable
- The plan cost is \$70 per month, charged month-to-month
- Modem charge is a fixed cost of \$252 if customers stay for less than 36 months, and free if they stay for 36 months or more
- The number of months, $x$, has a Gamma distribution with shape $\alpha = 6$ and scale $\frac{1}{\beta} = 6$. This is a distribution with a theoretical mean

  $$\mathbb{E}\left[x_{\text{Gamma}}\right] = \frac{\alpha}{\beta} = 36$$

  to reflect the fact that a customer would most likely want to stay in a plan for 3 years so that they don't have to pay for the modem upfront. It has longer tails than a Poisson distribution, which accommodates for those that may stay for less than 12 months, or much more than 36 months.

## B Cost sensitivity analysis details

The varying quantities for this sensitivity analysis will be cost-related, with the understanding that reduced revenues from customer dissatisfaction are also considered a cost:

- Telephone calling rate: 10c to 30c per minute
- Cellular calling rate: 25c to 75c per minute
- Gamma scale parameter $\frac{1}{\beta}$, lower meaning that customers stay for fewer months on average because of their disappointment with the level of services offered: 6 to 0.1

These ranges represent different market conditions, where extreme conditions are also included (i.e. costs grow threefold and acquired customers only stay for half a month on average).

# C HTML file for the 3d sensitivity analysis plot

Here is a Google Drive link to the exported 3d sensitivity analysis plot, which is a HTML text file that you may download locally and open in a browser. Actions that can be performed on this plot are located on the top-right hand corner:

- 360-degree rotating (click and drag)
- Panning and zooming (click and drag)
- Taking static screenshots and downloading as png

You can also hover on individual data points for more details on each constraint and the resultant price difference.