PREDICTING LISBON HOUSE PIRCES WITH TREE-BASED MODELS

A PREPRINT

Julia Przybytniowska
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, ul. Koszykowa 75
j.przybytniowska@gmail.com

Hubert Ruczyński
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, ul. Koszykowa 75
hruczynski21@interia.pl

Kacper Skonieczka
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, ul. Koszykowa 75
k.skonieczka01@gmail.com

May 11, 2022

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords Tree-based models · Regression · Lisbon House Prices · XAI

1 Data

Lisbon House Prices is a rater small data set from Kaggle. It contains 250 observations, which represent estates for sale in Lisbon, which are technically described by 17 columns, however some of them are unnecessary. In the end we decided to remove *Id* column which gives us no information at all and columns *Country*, *District* and *Municipality*, because all records have the same values there. Despite this columns the data set doesn't have any corrupted records and NULL values. Additionally we provide the description of every variable in the data set:

- Id: is a unique identifying number assigned to each house.
- Condition: The house condition (i.e., New, Used, As New, For Refurbishment).
- PropertyType: Property type (i.e., Home, Single habitation)
- PropertySubType: Property Sub Type (i.e., Apartment, duplex, etc.)
- Bedrooms: Number of Bedrooms
- Bathrooms: Number of Bathrooms
- AreaNet: Net area of the house
- AreaGross: Gross area of the house

- Parking: Number of parking places
- Latitude: Geographical Latitude
- Longitude: Geographical Longitude
- Country: Country where the house is located
- District: District where the house is located
- Municipality: Municipality where the house is located
- Parish: Parish where the house is located
- Price Sq. M.: Price per m² in the location of the house
- Price: This is our training variable and target. It is the home price

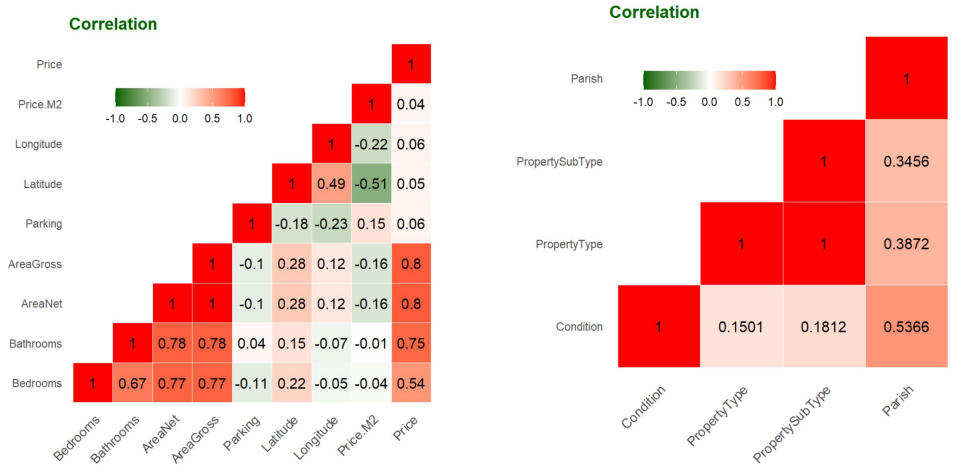
1.1 EDA

1.1.1 Dependance of numeric variables

To know more about the regression task that we are facing, we decided to conduct an Exploratory Data Analysis. We've created correlation heatmaps for numeric and categorical variables separately. Thanks to this method we were able to see the high dependence indexes between Bathrooms and Area Net, Bathrooms and Area Gross and most importantly, an indistinguishable differences between AreaNet and AreaGross. The latter ones give us the same information (correlation index equal 1), so we decided to truncate one of these columns for training process.

1.1.2 Dependence of categorical variables

Set of categorical variables is much smaller and consists of 4 variables only, however, we've manage to see another important dependence in the data set. Creation of Crammers V Correlation Heatmap showed us a strict dependence between PropertyType and PropertySubType variables. Deeper analysis showed us that it is caused by absolute dominance of Apartment PropertySubType.

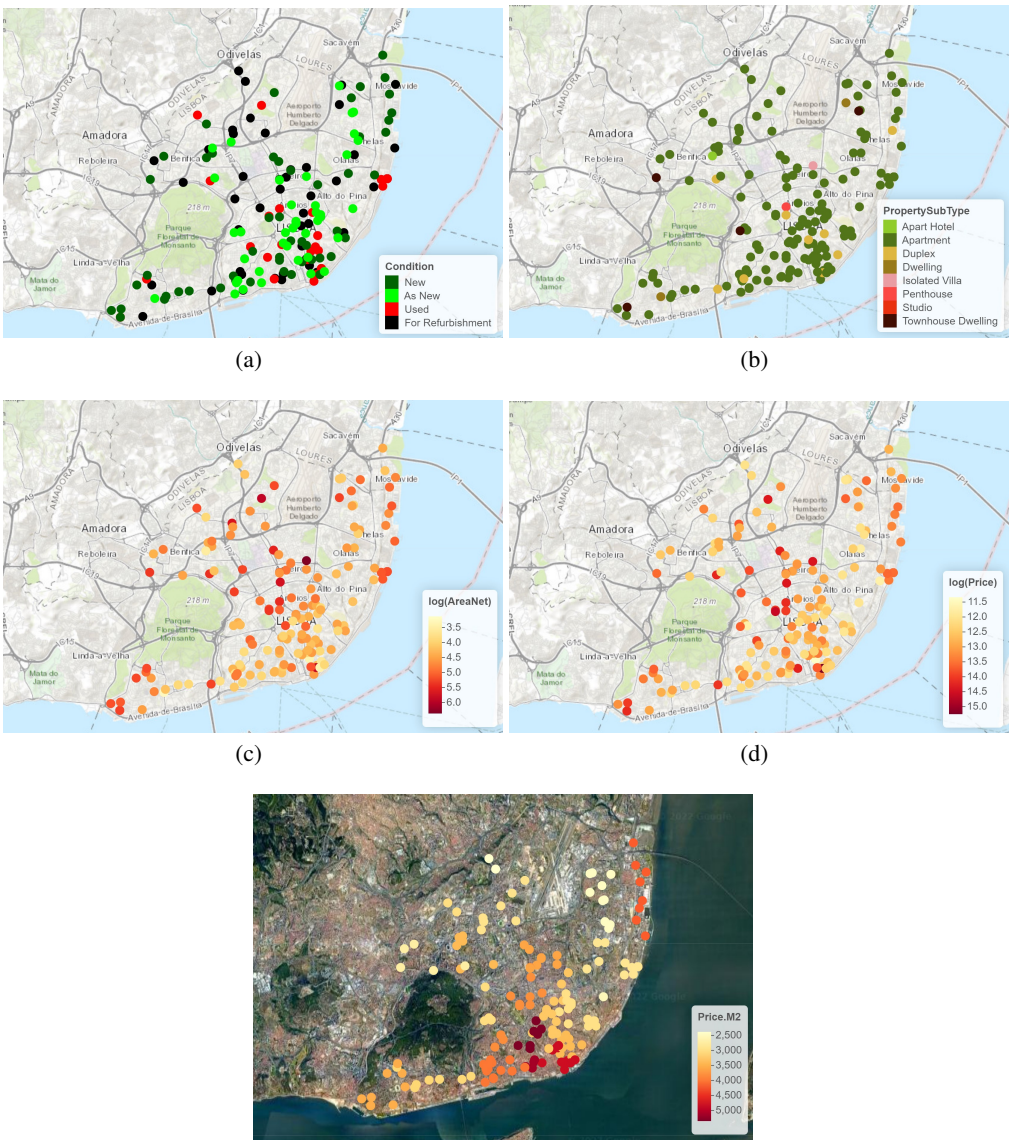


(a) (b)

Figure 1: Correlation Heatmap of: (a) numerical variables (b) categorical variables

1.1.3 Map visualization

To get more information about estate market in Lisbon, we've created maps of the estates presented in the data set. Fist thing we saw is the aforementioned Apartment dominance of the market. Moreover there is a high similarity of distribution between AreaNet and Price of the property. Apparently the most expensive properties are not located in the city center, despite the fact, that the highest prices per squared meter are obviously in the city center.



(a) (b) (c) (d) (e)

Figure 2: (a) Property Condition (b) Property SubType (c) Property AreaNet (d) Property Price (e) Property Price m^2

2 Models

In this section, we will present hyperparameter training method used by the authors, models trained with aforementioned parameters portfolio, compare and diagnose them to finally establish which models are the best.

2.1 Hyperparameters Optimization

In order to find best models we used Bayesian Optimisation method from 'ParBayesianOptimization' library. In each model our grid was a little different, what is presented in the table below.

2.2 Best models selection

Our selection method consists of three main steps:

- RMSE and MAE test values comparison

Model	Random Forest	XGBoost	LightGBM	CatBoost
Grid	num_trees mtry min_node_size max_depth	eta max_depth min_child_weight subsample lambda alpha nrounds	max_depth min_data_in_leaf num_leaves bagging_fraction	depth n_estimators l2_leaf_reg min_data_in_leaf learning_rate

Table 1: Hyperparameters' grid.

- RMSE and MAE train vs test analysis
- Split Groups plots analysis

Exact values presenting RMSE and MAE are included inside Table 2, however the easiest method to compare them is via plots present on Figures 3 and 4. These visualizations plots RMSE or MAE train values against equivalent test value in logarithmic scale. This way, we can detect models that perform the best for each metric or we can find models which are overfitted (far from the y=x line).

Model Name	RMSE train	RMSE test	MAE train	MAE test
Random Forest 1	110807	150676	50742	107078
Random Forest 2	118840	144766	57910	104303
Random Forest 3	128724	136672	57366	100909
Random Forest 4	459960	307582	311957	248038
Random Forest 5	172064	139157	76065	103170
xgboost 1	74917	280031	38016	94336
xgboost 2	187091	377060	104924	105870
xgboost 3	115302	354484	54714	110332
xgboost 4	13005	285468	10056	98338
xgboost 5	211260	386178	111113	112593
lightgbm 1	80114	295373	60760	112940
lightgbm 2	75805	285762	57978	115679
lightgbm 3	30010	302979	24968	118603
lightgbm 4	105949	374795	61276	143829
lightgbm 5	86672	336088	65598	129747
catboost 1	21717	193699	17465	116533
catboost 2	57757	212392	45237	122890
catboost 3	463355	337042	303518	264884
catboost 4	418607	306467	264996	235761
catboost 5	455769	331211	296353	259342

Table 2: Comparison of all 20 trained models

Accordingly to this selection method we narrowed our set of 20 models to just 4, which are presented in the Table 3. The pros of every model are presented below:

- Random Forest 3 is not overfitted and best performing model in terms of RMSE
- Random Forest 2 is second best model in terms of RMSE
- xgboost 1 has the lowest MAE test metric, however it is overfitted
- xgboost 2 is not overfitted in terms of MAE and has one of the best MAE scores

In the end, we created Split Groups plots to see, how our models classify observations. These plots automatically create bins for 8 subgroups of price ranges of the original data set and saves where specific observation lands. Later they take predicted values and do the same thing for them. In the end they take these observation and compares in which bin they were originally and where did they land after the prediction. The outcomes for our models are presented on Figure 5.

According to this diagnostic we decided to choose 2 models for XAI. First is **Random Forest 2**, it predicts well value of the cheapest houses and doesn't mistake them with houses that price is in 5 bin what Random Forest 3 does. In fact it

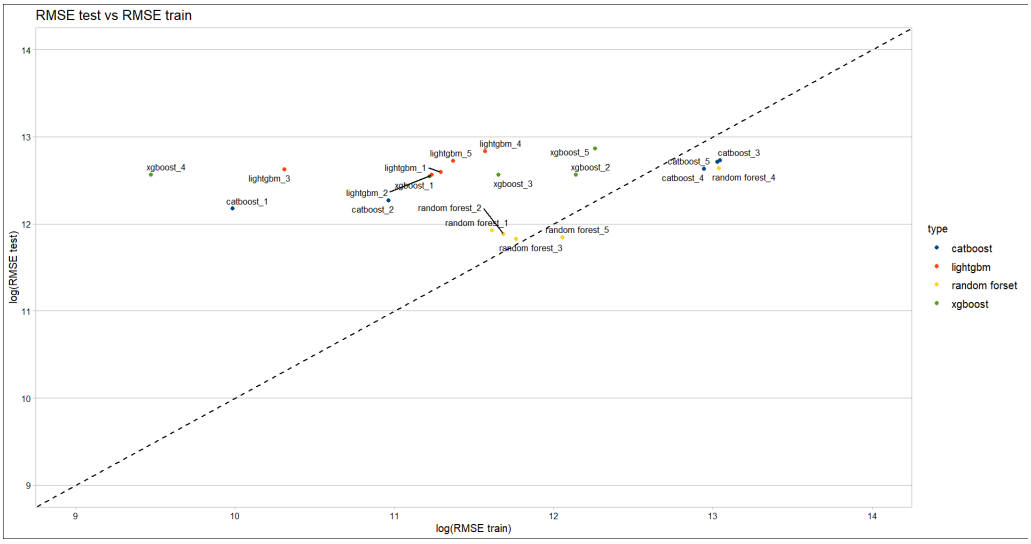


Figure 3: Train and test RMSE values comparison for all trained models in logarithmic scale

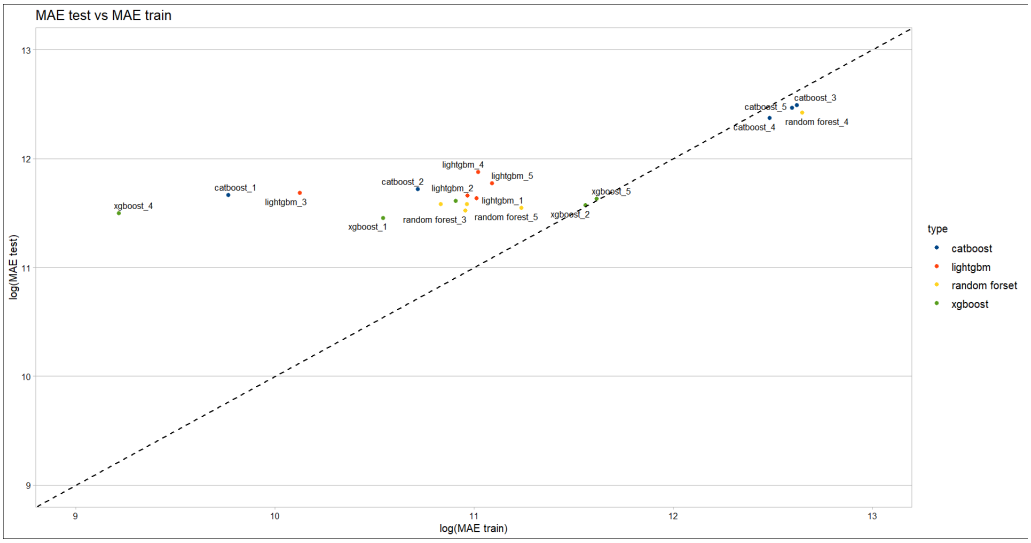


Figure 4: Train and test MAE values comparison for all trained models in logarithmic scale

Model Name	RMSE train	RMSE test	MAE train	MAE test
Random Forest 2	118840	144766	57910	104303
Random Forest 3	128724	136672	57366	100909
xgboost 1	74917	280031	38016	94336
xgboost 2	187091	377060	104924	105870

Table 3: Preselected models

predicts price in this bin for the most expensive houses but their bin is closer to it than the first one. Second chosen model is XGBoost 2. It has one of the best MAE scores and it distinguishes simultaneously the cheapest and the most expensive houses well, where mistake would be potentially the most damaging.

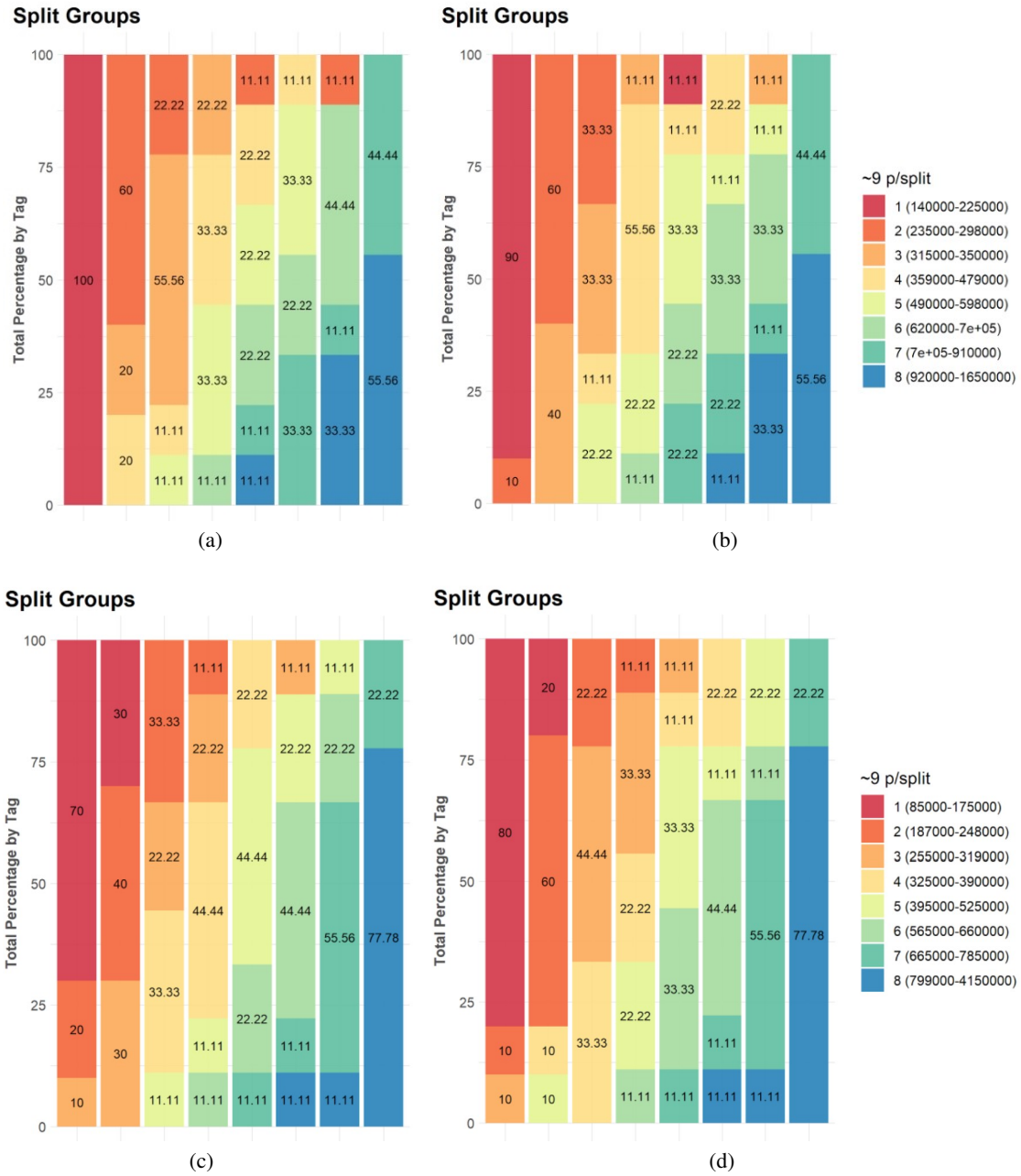


Figure 5: (a) Random Forest 2 (b) Random Forest 3 (c) XGBoost 1 (d) XGBoost 2

3 Introduction

4 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 4.

4.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

4.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

5 Examples of citations, figures, tables, references

5.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [Kour and Saabne, 2014a, Hadash et al., 2018] but other people thought something else [Kour and Saabne, 2014b]. Many people have speculated that if we knew exactly why Kour and Saabne [2014b] thought this...

5.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 6. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

5.3 Tables

See awesome Table 4.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

[1] Sample of the first footnote.

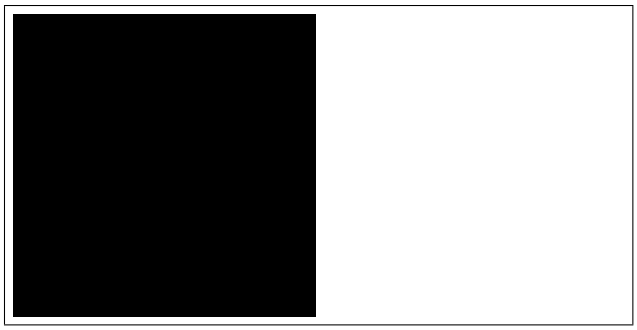


Figure 6: Sample figure caption.

Table 4: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

5.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014a.

Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014b. doi:10.1109/SOCPAR.2014.7008025.