Development report and deployment

Rotterdam Accommodation Price Predictor (RAPP)

Prepared by Pham Nguyen An Phuong

Date: 09/05/2021

Contact: 426122@student.fontys.nl

Abstract

The document contains process report, contextual information, and deployment information as well as future steps for the project Rotterdam Accommodation Price Predictor (RAPP). The aim is to predict the rental prices of accommodations in Rotterdam using infrastructure facts (e.g. surface area, number of rooms) and convenience factors (e.g. number of supermarkets in the area). Details regarding potential impacts have also been provided. Overall, RAPP has reached the suitable status regarding the time and content constraint of the project. Although the result can still be improved before official deployment, wireframe and infographics could be made. Furthermore, it is also possible to create prototype using the current algorithm and other languages/python libraries/tools; however, this is not within the scope of the current stage. For better result, looping through the second and third phase of the AI methodology is needed to gather more features. This would be another iteration and is not within the constraint of the current project. However, the scope can be extended in the future if there is a suitable timeframe.

Contents

1. Context....1
2. Process report....2
   2.1. General....2
   2.2. Proposal, exploration, data acquisition....2
   2.3. Modelling....2
3. Deployment....4
4. Conclusion....4
5. Appendix....4

1. Context

This is the 'Development report and deployment' document for the project Rotterdam Accommodation Price Predictor (RAPP). It contains the background, process, result report as well as the future path to be taken. Regarding RAPP, it is a machine learning project that aims at predicting the rental prices of accommodations in Rotterdam using infrastructure facts (e.g. surface area, number of rooms) and convenience factors (e.g. number of supermarkets in the area). This can be achieved by scraping data from one of the most used rental websites in the Netherlands - Pararius and applying cleaning – modelling via Python. There are five main stake holders including:

Stakeholder	Description
Vietnamese students in Rotterdam	These stakeholders are the main target group and inspirator for the development of RAPP. The developer hopes that the solution can help them in estimating rental price based on desired inputs (e.g. desired living area). They have moderate influence on the development process, but high interest.
General community in Rotterdam (Vietnamese and other professionals, and students of other nationality)	Although they are not the main target group, the solution is inclusive for all. Similar functions can be performed, with interest and influence being alike to the previous group.
Rental agencies and landlord in Rotterdam	By applying the solution, potential tenants will have already realized the suitable price range for what this stakeholder group offer. Hence the mitigation of bias complaints and waiting time for potential tenants to make a decision. Furthermore, this group may use the solution to check their own offered prices. They have low interest and low influence.
Data providers	Data is needed to carry out this project. Hence, providers (passive or not) are very important. The current provider for the pre-anchored data is Pararius. This group have low interest and high influence.
Real estate regulation experts	Beside complying with data regulations, no real estate law should be broken. Hence, consultation with experts in the field may be needed at some point of the project. They have low interest and possibly high influence.

2. Process report

2.1. General

The project was initiated since 19/04/2021 with a time constraint of 3 weeks for one cycle through the AI methodology, with four phases including proposal, provisioning, modelling, and deployment. All deliverables that have been handed-in are as follow:

- Proposal
- Scraper
- EDA and Provisioning notebook
- Modelling notebook
- Data requirement and collection ledgers
- Process report and demonstration material

The framework of the mentioned methodology can be found in the appendix. For version control, the related notebooks and documents can be found on GitHub. Considering the current status, it is concludable that the first cycle of development for RAPP has been completed on schedule. The result is currently in the form of machine learning models (XGBoost Regressor) that predicts the rental price in euros per month by taking the most prominent features in the dataset including:

- Surface: the surface area of the accommodation
- Distance: km to the selected point in the center (see EDA and Provisioning notebook for specific detail)
- rooms: number of rooms in the accommodation
- num_service: number of services in the area
- sp_num: number of supermarkets in the area

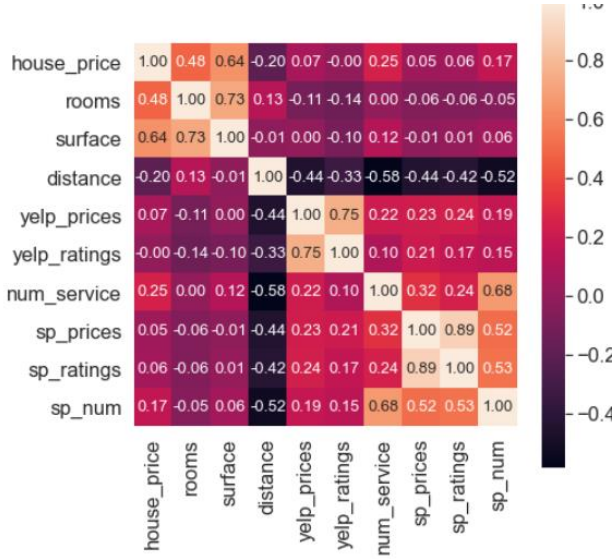
2.2. Proposal, exploration, data acquisition

The proposal with a full impact analysis using the TICT tool have been provided. Overall, no major concern that could serve as a setback to the direction RAPP is heading in could be found. This is because no communication or transaction is performed when using the product; hence there should be no opportunity to perform thieving or scamming for valuable goods. Furthermore, the collection and application of data within this project is straightforward, independent of personal issues of any party.

Regarding data exploration and further collection, a scraper has been built in Python to extract data into tabular format from html bins on Pararius. The result is 1428 rows of data with 7 columns related to accommodations in Rotterdam. The following step is adding convenience factors. There are three of them anchored: distance to Blaak (in city center), average number and ratings of supermarkets in the area, and similar information for services (cafes, bars, restaurants). These are calculated and joined using coordinates and postcodes. To keep track of the data to be scraped, the Data requirement and Data collection ledger has been used.

2.3. Modelling

The modelling process can be carried out using the data file from the previous phase. The first action to take is feature selection using heatmaps.



Above is a heatmap which displays Pearson's correlation coefficient - the test statistics that measures the association between two (numerical) variables. From it, information about the magnitudes of the relationships can be derived. There are five features which has noticeable correlation to the target ‘house_price’ and was selected for modelling (as mentioned in 2.1).

The first model used is Random Forest Regressor. This is a versatile ensemble technique with various parameters to aid in tracking and prevention of overfitting. The essential working method behind this model is to create different trees with bootstrapped samples of the training data, train the model for each sample, then average the predictions to get a final result.

The second model is XGBoost Regressor – another ensemble type algorithm. Essentially, the idea to it is learning from past results to enhance predictions. There are three elements to this type of model, including a loss function to be optimized (e.g. squared error for regression), a weak learner to make predictions (short and simple decision trees), and additive modelling to add weak learners that can do what their predecessors could not.

The final model tried is SVR. It adds a tube around the estimated function, which width is determined by a parameter called ε (epsilon). All points that fall within this tube are considered as correct predictions. Although based on SVM, the support vectors for SVR can also be the points that fall outside the tube rather than just the ones at the margin. The distance to these points is controlled by 'slack' or the C parameter.

After modelling the baseline version, all models received parameter tuning to deal with overfitting and maximize performance. To analyze the results, R^2 and RMSE is used. The former measures the proportion of the variance for a dependent variable (y) that can be explained by independent variables (X) in a regression model. As for RMSE (Root Mean Square Error), it signifies the average difference between the observed known values of the outcome and the predicted value by the model (same unit as the dependent variable). Overall, a high R^2 and low RMSE is desirable.

The final result indicated that the most suitable model is XGBoost since it has the highest R^2 and lowest RMSE out of all the tuned models. More specifically, 52% of variation in rental price of testing data can be explained by the selected features and the root mean square error is 623.834 if XGBoost is used. Nonetheless, there is still room for improvement as the current

RMSE is rather noticeable considering the distribution and unit of the rental prices (more detail in the modelling notebook). Given the time constraint in the proposal, the path ahead will be to report on the findings, prepare demonstration material, and return to stage 2 and 3 at a more available time the future.

3. Deployment

The vision for RAPP is a website that targeted users can log on, input required features, and get helpful insights. This means that the deployment of the product will have to deal with web coding. There exist different methods for this, including Heroku, coding with django, or using tkinter. Once the front-end design is completed, a hosting server will need to be bought in order to make the website fully functional. Further deployment could be the embedding of the product to the association of Vietnamese Student in the Netherland's website. There should also be disclaimer notes included on the site regarding relativity of prediction, as well as instruction for usage.

Although it is not within the scope of this project to deploy the product itself, demonstration material will be created as specified in the deliverables. This will be in the form of infographic to explain functionality and wireframe to show the envisioned front-end.

4. Conclusion

From the current result, it is concludable that RAPP has reached the stage that it should. Although the current result can still be improved before official deployment, wireframe and infographics could be made. Furthermore, it is also possible to create prototype using the current algorithm and other languages/python libraries/tools; however, this is not within the scope of the current stage. For better result, looping through the provisioning and modeling/evaluation stage of the AI methodology is needed. Specially to gain more resourceful features. It is noteworthy that the current data storage is local machine since it is convenient and simple to perform with respect to the data size.

5. Appendix

Below is the aforementioned AI methodology framework: