Big Data Management and Analytics - Group Project

November 15, 2022

INSTRUCTIONS

- This is a team assignment. Teams must have 4 or 5 students and all team members will receive the same grade.
- Submit your proposal digitally as a single **PDF file** via Canvas. The file name should follow the convention: `Project_Team_XX.pdf` in which `XX` corresponds to the two digits that identify your team.

Deadline:

Friday, December 2, 2022 at 23:59.

Late submissions are not allowed

Warning: The detection of any form of plagiarism in your work means the assignment will be graded with ZERO points.

RealtyCheck

Your second project at Double Zero Data, the prestigious analytics consulting firm you work for, is to help RealtyCheck, a real estate firm, in making smarter investments. RealtyCheck focuses on acquiring and selling houses/apartments in Rotterdam, taking advantage of market fluctuations. In some cases RealtyCheck adds value by making renovations to the property before listing them for sale, in other cases they simply identify good opportunities and perform a quick sell at a profit.

Part I (6 points)

RealtyCheck hired Double Zero Data to help them develop an automated system to identify attractive real estate properties as they come to the market before their competitors and to advise on their value. Your job is to propose a full-fledged solution for RealtyCheck, covering all steps of the CRISP-DM cycle. Your proposal should address at least the following concerns expressed by RealtyCheck:

1. What is the potential added-value that such a system will bring to RealtyCheck? What are the strategic advantages?
2. What kind of metric should be considered when assessing whether a specific property is attractive, considering RealtyCheck's business model?
3. What type of model should be developed (classification / regression / clustering, etc.)?
4. What kind of information should be collected, from which sources, and how often in order to assess the attractiveness of a new property in the market?
5. How can we assess the performance of such a system when in production? How often should it be updated?

You should also address any other questions you find important but that were not explicitly raised by RealtyCheck. Your proposal should be persuasive and demonstrate the value of your solution, but also honest about its potential weaknesses.

- Note: One strategy that may be useful to approach this problem is to start by assuming you have access to all the necessary resources to implement the ideal solution. As a second step you may add some plausible restrictions and then try to work around them.
- Note: You do not need to mention all the concepts, models and tools covered in class just because. If a specific model, concept of tool does not apply or is not adequate to this specific case, please do not recommend it to RealtyCheck.

Part II (4 points)

As a first step towards a complete solution, you decide to assess how easy it is to predict house price categories as a function of some of their publicly available characteristics. This task will provide

you with some information on the likelihood that you can develop a good solution for RealtyCheck. After some internal conversations, you find out that a former co-worker has already scraped a real estate web site and gathered information about all properties available for sale in Rotterdam. After some data cleaning they were able to assemble a data set with information that may (or may not) be useful to predict the listing price of a property, including location (zipcode), area, amenities, along with other information.

You should use these data to run the models you believe are adequate and that you believe will provide the best predictions. You can use the knowledge acquired in the Individual Assignment as starting point for this task. This task is different from the task in the Individual Assignment in at least one important aspect. There are more than two classes. Your goal is to predict to which price range (out of 5 ranges) does a house belong to. Class ranges (or categories) are:

Label	Meaning
"< 100K"	Houses with price below 100,000 Eur.
"100K-150K"	Houses with price above or equal to 100,000 Eur. and below 150,000 Eur.
"150K-200K"	Houses with price above or equal to 150,000 Eur. and below 200,000 Eur.
"200K-300K"	Houses with price above or equal to 200,000 Eur. and below 300,000 Eur.
"> 300K"	Houses with price above or equal to 300,000 Eur.

You should reflect on your results (what do they mean) and on potential uses of this model taking into account RealtyCheck's business model. Specifically, you should answer the following questions:

- What is the meaning of performance as it is measured right now?
- Which variables contribute the most for model performance?
- Would such a model be useful for RealtyCheck? If so, under which conditions/assumptions?

Notes:

- There are no limits at all to which models and techniques you can use, as long as you show you understand the concepts used. We suggest that you first try the simpler models, such as decision trees, and only afterwards explore more advanced techniques, such as bagging and boosting.
- Although this task may be informative about the potential success of your solution, your proposal (in Part I) does not need to be limited to the data used and results obtained in this task. For Part I you can assume you will have access to other (potentially more interesting) data sources as well.

Kaggle Competition

To make this task more interesting, you will be competing against other teams to develop the best model. In the context of this competition, the best model is the model that predicts outcomes with highest accuracy, so you should use the variable `price_category` as your target variable.

The competition is hosted by Kaggle, specifically at:

- `https://www.kaggle.com/competitions/realtycheck-bdma2022`

Note that to participate in this project you will need to use the following link:

- `https://www.kaggle.com/t/caf6526d85574938a7f1cc0fcdcf568a`

In the competition you will have access to two files:

- `dt_realestate_train_2022.csv` should be used as your training set;
- `dt_realestate_test_2022.csv` should be used as the test set.

In order to prevent overfitting, the test set does not contain the `price_category` attribute. To measure the performance of your model in the test set, you need to prepare a solution file and submit it on Kaggle. Kaggle will then report back your performance and rank you against other teams. You can prepare the solution file by making use of the `Write CSV` operator together with `Select Attributes` operator to automatically create a file with only two columns: `id` and `prediction(price_category)` (your prediction).

- You may need to **activate the option "Include special attributes"** in the `Select Attributes` operator, so that the generated CSV file contains only two columns (we believe this is a glitch in RapidMiner).
- It may help that you define your `id` variable to be a String when importing the data (import wizard), otherwise you might get ids such as *3342.0*, which will give you an error when submitting to Kaggle. See the file `dt_realestate_sample_solution_2022.csv` for an example on how the file should be formatted.

For us to track your progress in Kaggle, please register your team as **BDMA2022-XX** where XX is the number of your group on Canvas.

Awards:

- The **top 10 teams in the final rank will be awarded one extra point** in the final grade of the project.
- By the end of the competition, we will ask the winners to share a brief description of their strategy & methods, so that all of us can learn from them. **This is a requirement for the award of the extra point**. You can share (part of) your write-up for the second part of the project.

Important Note:

- By the end of the competition you will be asked to choose three of your submissions to be considered for the competition. By default, the three best submissions in the public score will be chosen. We suggest that you choose the submissions you believe are the most robust ones, even though they may not be the ones scoring at the top in the public score. Note that the private score is calculated from data that you have not seen before, so your best solutions in the public score may be overfitting the respective test set. It frequently happens that only a few of the best 10 teams in the public score remain in the top 10 in the private score.

The Data

The original data were scraped from an online platform using the `webscraper.io` Chrome extension mentioned in class.

You will be using cleaned-up data based on the original scraped dataset with some (but not all) of the information available from the website. The table below briefly describes each of the available fields.

VARIABLE	ROLE	TYPE	DESCRIPTION
`id`	identifier	numeric	Unique property identifier
`price_category`	target	multinomial	Price range; 5 categories; see table above
`year`	predictor	numeric	Construction year
`zipcode`	predictor	multinomial	ZIP code of the property
`total_rooms`	predictor	numeric	Total number of rooms
`bedrooms`	predictor	numeric	Number of bedrooms
`n_weeks_old`	predictor	numeric	For how many weeks has the property been listed
`living_area`	predictor	numeric	Total living area (in square meters)
`other_area`	predictor	numeric	Other areas (in square meters)
`total_area`	predictor	numeric	Total area (in square meters)
`monthly_contrib`	predictor	numeric	Mandatory monthly contribution to owners association
`n_photos`	predictor	numeric	Number of photos posted
`n_photos_360`	predictor	numeric	Number of 3d photos
`type_of_construction`	predictor	multinomial	Type of construction
`other_indoor_space`	predictor	numeric	Other indoor areas (in square meters)
`energy_label`	predictor	multinomial	Energy label
`located_on`	predictor	numeric	Floor number
`own_ground`	predictor	binomial	Whether the property is located in own ground
`flg_missing_year`	predictor	binomial	Whether the variable `year` was missing
`flg_missing_total_rooms`	predictor	binomial	Whether the variable `total_rooms` was missing
`flg_missing_bedrooms`	predictor	binomial	Whether the variable `bedrooms` was missing
`flg_missing_n_weeks_old`	predictor	binomial	Whether the variable `n_weeks_old` was missing
`flg_missing_living_area`	predictor	binomial	Whether the variable `living_area` was missing
`flg_missing_other_area`	predictor	binomial	Whether the variable `other_area` was missing
`flg_missing_total_area`	predictor	binomial	Whether the variable `total_area` was missing
`flg_missing_monthly_contrib`	predictor	binomial	Whether the variable `monthly_contrib` was missing
`flg_missing_n_photos`	predictor	binomial	Whether the variable `n_photos` was missing
`flg_missing_n_photos_360`	predictor	binomial	Whether the variable `n_photos_360` was missing
`flg_missing_type_of_construction`	predictor	binomial	Whether the variable `type_of_construction` was missing
`flg_missing_other_indoor_space`	predictor	binomial	Whether the variable `other_indoor_space` was missing
`flg_missing_located_on`	predictor	binomial	Whether the variable `located_on` was missing
`flg_missing_own_ground`	predictor	binomial	Whether the variable `own_ground` was missing

Format

Your proposal will have a title page with your names, student numbers and team numbers, followed by an introductory text explaining the context and the business problem. After this you will have up to 3,000 words (that must fit in 5 pages, including tables and figures) dedicated to Part I, and up to 3,000 words (that must fit in 5 pages, including tables and figures) dedicated to Part II. Note that Part I should be your proposal for RealtyCheck to be read by their executives, while Part II can be interpreted more of as an internal technical support document to be used by your fellow colleagues in case the proposal is accepted. References do not count to the page limit.

Good Luck!