Data Mining Assignment - Classification

Sam Pinxteren, Prof. Calders

Deadline: April 27, 23:59

This assignment document consists of 3 sections; the project description, the tasks and questions, and the notices. Please read all carefully before beginning your work.

1 Project Description

Imagine the following scenario; in a large city, the housing market is very controlled. Houses available for sale are listed in an online database. Real-estate agents can access this database and choose to commit to selling a certain house, making this commitment costs €450. The profit from selling one of the houses on the list is based on whether the house has a high or a low value (we consider this a binary attribute in this case). Selling a high value house nets an income of €600, selling a low value house earns €100. This means, if you commit to selling a low value house, it costs you €350.

There is a catch however, there is already another real-estate agent taking up these contracts. If you and the existing real-estate agent were to both take up a contract, it costs you both €450, but the earnings will be split. This means that if you both take up a contract, you both lose money. You probably want to avoid taking any contracts the other agency takes up.

All knowledge about the houses and sales is public. So you get information about the available properties, and you can see which houses were selected by the existing company. This means you might be able to predict which houses are high value and which houses the other company will not select to sell.

This assignment comes with two CSV files *historical.csv* and *current.csv*, in which each line corresponds to a single house. The historical dataset describes the houses that were available in the past, including their value status and whether the existing company has taken the contract. You will use this historical dataset to make predictions about the *current* dataset. The latter contains houses for which the contracts are available now.

- *identifier*: A unique identifier for each property
- *size*: The size of the property.
- *kitchens, bathrooms & rooms*: The number of each type of room in the property
- *floor*: The floor the property is on
- *type*: A categorical value indicating the type of the building
- *year*: The year in which the building was constructed
- *condition*: A categorical value which says something about the condition
- *elevator*: A boolean value indicating the presence of an elevator

- *subway*: A boolean value indicating whether a subway station is close
- *district*: A categorical value indicating the district
- *recentOwner*: When true, the previous owner had this property for more than 5 years
- *longitude & latitude*: The position of the property

The train dataset has two additional columns:

- *highValue*: True when the property is high value
- *prediction*: True when the existing real estate agency has taken this contract

2 Tasks and Questions

Complete the following tasks and report your process together with the answers to each question in a pdf.

1. In the historical dataset (*historical.csv*), how did the existing company do? How much did they earn, assuming taking a contract also costs them €450?
2. What percentage of the high value properties did the company take the contract on, why do you think this is not higher?
3. The file *current.csv* contains the houses with open contracts. Here you do not know which houses have a high value or which contracts the other company will take. You may assume the other company will use the same model they used on the historical data to make their decisions, however. How much do you think the company would earn if they maintained their monopoly?
4. Implement some way to decide which contracts to take to maximize your profit. Create a file which contains all the identifiers of the houses for which you want the contract. One identifier per line in a file called *selection.csv*.
5. How much do you expect to earn for the houses in your submission?

3 Notices

Your final submission should be a zip-archive called submission.zip which contains:

1. Your report, a PDF file called *report.pdf*. Although it is not a requirement, we recommend you use LaTeX.
2. The code you have written to solve the questions. The report should stand on its own as your submission. The code should not be necessary for us to evaluate your work.
3. A file called *selection.csv*.

Important Notes:

- Any file with the wrong file name will not be considered, including capital letters and file extensions. Please double check that this is correct.
- Upload your submission to the assignments section on blackboard for this course and assignment before the announced deadline, any delay will result in a score of 0 for this assignment.
- If you have any questions or remarks, feel free to send an email to: sam.pinxteren@uantwerpen.be.