1 introduction

Imagine the following scenario. You want to start up a new real estate company in a large city. At the moment there is only one other company active in the city. The earnings of any company who wants to sell properties in this city are strictly controlled. The amount a company can earn by finding a buyer for a property is based on only one thing. If the property has a high value per square meter, the real estate agent gets €600. For a property with a low value per square meter, the agent gets €100.

The existing company saw an opportunity to earn money by using classification methods. They use an algorithm which takes in information about the properties and predicts if it is a low or high value property. Based on these predictions they try to optimize their profit by avoiding taking on sales for low value properties. They do this because all the costs in finding a buyer adds up to more than €100. So dealing with those properties makes them lose money.

All the information on the properties and sales is public. Now, you've kept an eye on this information and have seen an opportunity to earn some money. You have done all the calculations and found that it would cost you €450 to find a seller for a property, regardless of it's value. That means that if you take on a contract for a high value property you make a profit of €150. There is a catch however, if you and the existing company both take on a contract, it will still cost your company €450, but the profit will be divided between you and the other company. This means you would lose money regardless of the value of the property.

The file *train.csv* contains all the data you have gathered. It shows the characteristics you could gather on each property. Besides that it shows whether the other company took on the contract and whether it ended up being a high or a low value property. Here is an explanation of the features you have on each property:

- *identifier:* A unique identifier for each property.
- *size:* The size of the property.
- *kitchens:* The number of kitchens in the property.
- *bathrooms:* The number of bathrooms in the property.
- *floor:* The floor the property is on.
- *type:* The type of the property.
- *year:* The year in which the building was constructed.
- *condition:* The condition of the property.
- *elevator:* Whether the building has an elevator.
- *subway:* Whether a subway station is close to the building.
- *district:* The district in which the building lies.
- *rooms:* The number of rooms in the property.
- *recentOwner:* Whether the previous owner had the property for more than 5 years.
- *longitude:* The longitude of the building.
- *latitude:* The latitude of the building.
- **highValue:** Whether the property has a high value per square meter.
- **prediction:** The prediction made by the existing company. When True they took on the contract, when False they chose not to take it.

2 Questions

Write a report answering the following questions. It is encouraged to read all the questions before starting. Some later questions might influence how you would solve the earlier ones. Submit a zip file containing the following files:

- The file containing the list of identifiers (see question 4).
- All code written to solve the given tasks, do not include the given data files.
- Your report, a pdf with explicit anwers to all 5 questions. Pay attention to the description of your experimental setup.

Question 1: You can see the *prediction* column has far less True values than the *highValue* column. Why do you think the existing company took on so few contracts?

Question 2: Do you have any idea which type of classification model was used by the other company to make their decisions? Is there any way you could figure this out by using only the information in the *train.csv* file?

Question 3: The file *test.csv* contains all the information you have on the properties with open contracts. Here, it is not given which properties have a high value or which contracts will be taken on by the other company. You may assume the other company will keep using the exact same model to make their decisions. How much do you think the other company would earn if they maintained their monopoly, i.e., if you didn't provide any competition as you intend to? You may assume it also costs them €450 to take on a contract.

Question 4: Implement some way to decide which contracts you would take on to maximize your profit. With your final submission, add a file that contains the identifiers of these properties (the number in the *identifier* column). A simple text file where each line contains an identifier suffices. In your report clearly outline what you have tried and why, what worked and what didn't.
Note: For many of these classification and data mining challenges it can be useful to add more features and data from external sources. For this assignment however we discourage you to do this. It is a deep rabbit hole that we don't want to burden you with. For this reason only methods using the data provided in the two files are valid.

Question 5: How much do you expect to earn with the submission you made? How much do you think the other company will earn assuming it also costs them €450 to take on a contract?

3 Helpful examples

You are not required to use any specific language or library. However, python is a very efficient language for tasks like these. The *pandas* package for python can be used to easily read data files like the ones provided for this assignment. The *sklearn* package provides many different classification algorithms. When python is installed pip can be used to get the latest versions of these packages:

```
> pip install pandas sklearn
```

In python, the files can be loaded as a pandas dataframe:

```
import pandas as pd
train = pd.read_csv("train.csv", index_col="identifier")
test = pd.read_csv("test.csv", index_col="identifier")
```

The web site of sklearn provides detailed instructions on how to get started with the library:
`https://scikit-learn.org/stable/tutorial/basic/tutorial.html`