Data Mining Assignment - Pattern Mining

Nick Wils

April 27, 2022

Abstract

You can find the code of my project with the following link: github_link.
In my code I a class DataFrameSelector() which I found on the following website: website

1 Introduction

In this report I will evaluate an existing Real-estate company and calculate expected earnings if I would enter the Real-estate business. In this report I will use the following rules:

- Commitment costs = €450
- Selling a high value house = €600
- Selling a low value house = €100
- Multiple agencies selling the same house = split net income

Note: When going through the databases I found strange data. It looks like the columns "year" and "type" are switched. Furthermore, it looked like columns "bathrooms" and "floor" are switched as well. It would be very unlikely that most houses have more then 10 bathrooms and are on the 3th floor. After loading in the CSV-files I switched the column names as described above. The change has very little influence on the final result. It only changes how the data is preprocessed. So even if the change is wrong, it would not make a big difference.

2 Earnings existing company

In the history dataset the existing company bought 36 382 houses. From these 36 382 houses, 29 649 are high value and 6 733 have a low value. With this information I can calculate the following data:

- Net income = €18 462 700
- Expenses = €16 371 900
- Profit = €2 090 800

This means that the accuracy of the companies model is 81,49359573415426%.

3 Percentage of high value properties of existing company

In the history dataset were 48 008 high value houses. Of those the company took on 29649 houses. This is 61.758456923846026% of the market.

There are multiple possible reasons as to why this percentage is not higher. First of all their prediction model is not perfect. This matters because for every 2 low value house you take on you will have to sell 7 high value houses to break even. This means that making more mistakes has a very big impact. At the start the market is split up in about 48% high value houses and 52% low value houses. When the company is finished that split changes to about 29% high value and 71% low value. This means that the chance that the company would take one a low value house by mistake rises.

Another reason can be found by looking at the correlation matrix of the dataset. Here you can see that there is a strong correlation between "district" and "prediction". This points to that the company is probably not active in some districts.

4 Predicting the existing companies future earnings

To calculate the future earnings when the existing company keeps their monopoly, I tried trained a model on the history dataset. After the training and testing I used that model to predict which houses the company would take on in the current dataset.

In this assignment I used the following libraries:

- pandas: for the datasets
- numpy: for sklearn outputs
- sklearn: for models and metrics

The first thing I did was deciding which features were not useful. Because there is already a feature district, I didn't think the longitude and latitude had any value. The identifier, highvalue and prediction features also needed to be filtered out because they would make the model learn wrong connections. So I filtered the columns: "identifier", "longitude", "latitude", "highValue" and "prediction" out of my dataframe. I later tested multiple combinations of features and found that this was the best combination.

Secondly I preprocessed my data. Because some features were categorical and some numerical, I needed two different ways to preprocess data. I used OneHotEncoder for the categorical features and StandardScaler for numerical features that could benefit from scaling. The other features did not benefit from preprocessing These are the features I preprocessed:

- StandardScaler = kitchens, bathrooms and rooms
- OneHotEncoder = type, condition, elevator, subway, district and recentOwner

To streamline this process I make use of pipelines.

After preprocessing my data I chose my model. I chose for the DecisionTreeClassifier, I tried out multiple classifiers and found that this one preformed best. I trained this model with my preprocessed data, with the "prediction" column as the solution. For testing I used (5)k-fold cross validation, which gave me the following scores:

- f1-score = 0.9933590903466197
- recall = 0.9929085811665109
- accuracy = 0.99517
- precision = 0.9938100085284327

My model is now trained and ready to use. But before I can make my prediction I have to preprocess the current data. I do this by removing the same features and then using the same pipeline with the OneHotEncoder and StandardScaler that I fitted to my training data. After this I can predict for each property if the existing company would buy it or not.

Finally I can filter out the properties where my model predicted that the company would by. Together with the accuracy of high value houses that the company showed in the history dataset I can predict how much the company will earn.

My prediction is that the company will have a total profit of 424 450 euro by selling 7 383 houses. From those 7 383 houses there would be 6 017 high values and 1 366 low value.

Because the recall and precision of my model are not perfect, there will be some deviation. With my recall I can calculate that I will have about 52 false negatives. In the same way I can calculate 46 false positives with the precision. This would be the deviation on my prediction.

5 Which houses would I buy

Because I would lose money if I buy the same property as the existing company, I first filtered out all the properties that I predict the other company would buy. To predict which of the remaining properties would be high value I needed another model. I took the same steps as I did to make the other model, but this time I used the RidgeClassifier as model and the 'highValue' column as test solution. I still used the full history dataset for training. I only filtered out the properties bought by the other company in the current dataset.

I changed the model class because this gave me the best results, which are:

- f1-score = 0.7468966968230591
- recall = 0.7024870854857523
- accuracy = 0.77143
- precision = 0.7973001725809121

Now all I have to do is filter out the the predicted low value properties from the dataset where I already filtered out the predicted buys from the other company. You can find the properties I would buy back in selection.csv.

6 What do I expect to earn

In my prediction I found 2 925 houses. In the best case scenario those houses are only bought by me and are all high value. In that case I would have a profit of 438 750 euro.

Sadly this will probably not be the case as my precision is only 79,73001725809121%. So in the worst case about 20% of the properties are low value.

In this case I will have a profit of 142 250 euro, where I sell 2 332 high values and 593 low value properties.

But something else can go wrong. It is possible that I predicted the properties of the other company partly wrong. To adjust for this I can use the recall of that model. The worst case scenario here is that every property I got wrong, is a high value one, that I bought as well.

In this case I will have a profit of 409 600 euro, where I sell 2 872 high values properties alone and 53 high value properties together with the other company.

In the absolute worst case these 2 things happen at the same time.

In this worst case I will have a profit of 126 350 euro, where I sell 2 279 high values and 593 low value properties alone and 53 high value properties together with the other company.

So taking this all in consideration, I would have a profit between 126 350 euro and 438 750 euro.

In conclusion if I would have to take a guess, I would guess the average of the best- and worst case: which is 282 550 euro.