Unit 1 - Cost of Teacher and Test - Evaluation, Deployment And Monitorisation Of Models

Ole Magnus Laerum and Piotr Franc

Polytechnic University of Valencia (UPV), Spain

1 Introduction

This paper looks at Logistic Regression, Regression Tree and RuleFit. For the Logisic Regression the dataset from COMPAS is used, and for the Regression Tree and RuleFit the Bike-Sharing dataset is used. This task was written in Python.

2 Logistic Regression - Task 1

2.1 Method

To solve this task, the COMPAS dataset was first pre-processed. This was done by:

- Removing rows that has $days\_b\_screening\_arrest > 30$
- Removing rows that has $days\_b\_screening\_arrest < -30$
- Only keeping rows with $race == Caucasian$ or $race == African - American$
- Adding a $recidivist$ variable that is 1 if $decile\_score > 4$

Then, the variables sex and $race$ were processed to binary 1 or 0 variables. Finally, a simpler dataframe X was created. The variable Y was set to the new variable *recidivist*.

Furthermore, the model was fitted and trained using the *glm()* function. To be able to use python instead of R, the library *statsmodels* was used.

2.2 Results

Finally, the following listed parameters were found and listed in the table **??**. The plot for $odds_ratio$ vs variables is shown in figure 1.

Variable	**Coefficient**	**Odds Ratio**	**Std. Error**
const	1.3266	3.7683	0.1230
age	-0.0486	0.9526	0.0029
is_recid	1.1231	3.0743	0.0617
sex_Male	-0.0494	0.9518	0.0776
race_Caucasian	-0.7692	0.4634	0.0636

Table 1: Variables and features and parameters from task 1

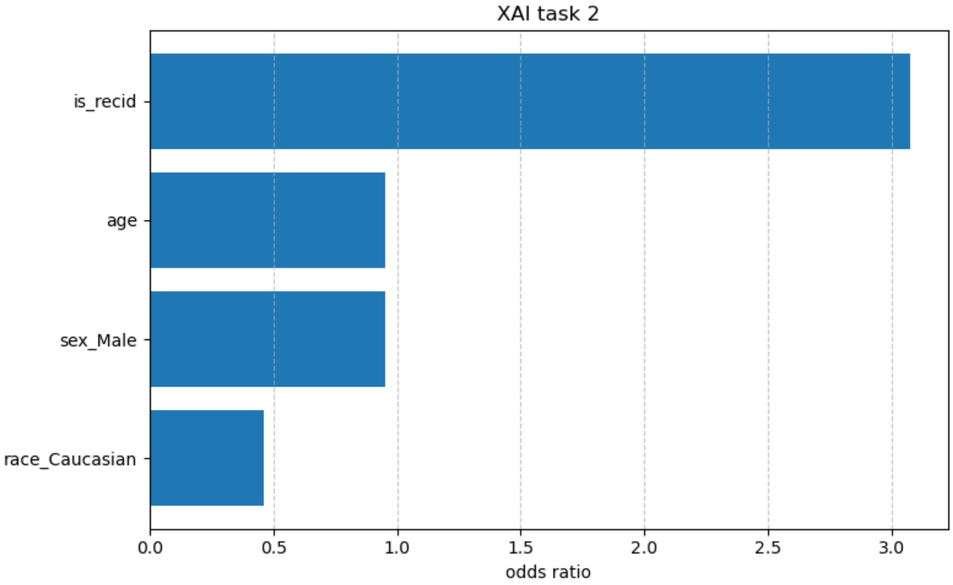

Fig. 1: Results from task 1

2.3 Discussion

As keen be visually seen from figure 1, *is_recid* is clearly the highest risk indicator of a person getting a high *decile_score*. The *odds_ratio* is 3.0743, which essentially means that people who actually have re-offended are **3x more likely** to be labeled a high risk person, that is, labeled as *recidivist* with a high *decile_score*.

Furthermore, race also has a significant importance, with a large negative coefficient of -0.7692 and a odds ratio of 0.46. This means, based on the COMPAS dataset, that being a Caucasion and not an African-American, means being **54%** less likely to me labeled a high risk person.

Age is to a small extent negatively associated with risk with a odds ratio of 0.95, meaning that each additional year gives a reduction of 5% odds of being labeled high risk.

For sex (being male), no statistical significanse was found given this preprocessing and selection of data.

3 Regression tree - Task 2

3.1 Method

In this task, the goal is to make a regression tree based on the *Bike-Sharing-Dataset* used in the XAI1 task.

Firstly, the dataset was prepared and preprocessed as done before, only that for this task, it is written in Python.

Then, a regression tree is built with a maximum depth of 2, using the *DecisionTreeRegressor* scikit library for Python, with the *cnt* variable as the target variable.

3.2 Results

The results are plotted using *matplotlib*'s *plot_tree* function. The plot of the tree is shown in figure 2.

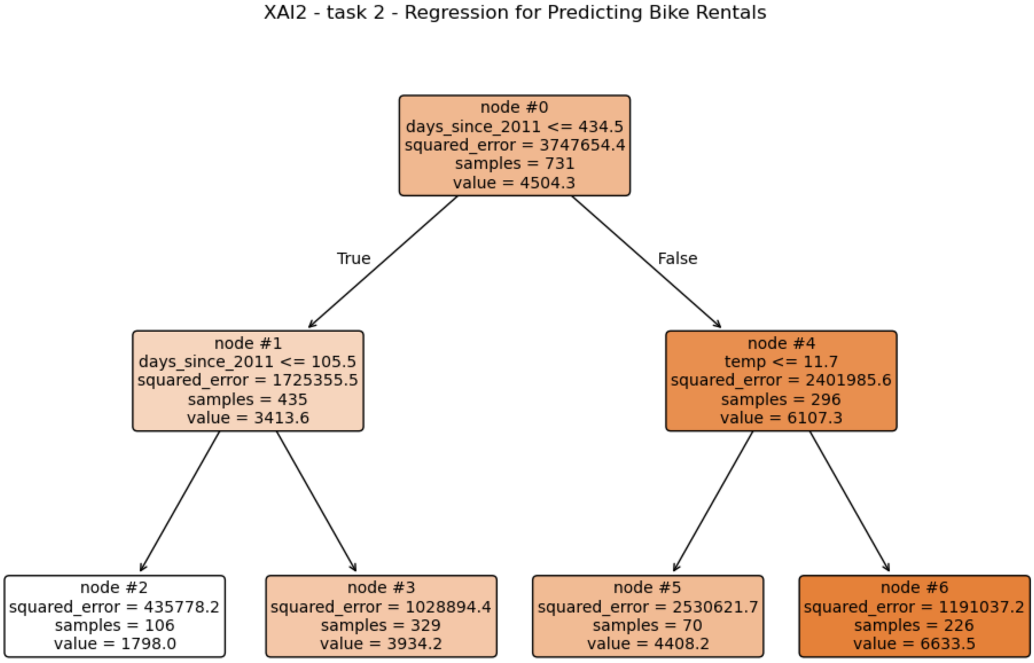

Fig. 2: Regression tree for task 2

3.3 Discussion

This regression tree "only" has a depth of two, which makes the accuracy not too good, but on the other hands, gives good interpretability, which is the main theme of this unit.

The first decision is made as **days_since_2011** ≤ 434.5. This means a cutoff of 10th March 2012. If the date is before this, it goes to **node #1**, or if it's after, it goes to **node #4**.

The second decision in **node #1** also depends on date, after or before 23rd June 2012. The third decision is based on temperature instead. Essentially, the decision in **node #4** says: *Given "late period", is the temperature bigger than 11.7 degrees?*

The **value** parameter shows the mean value of the bikes rented in that period. Based on the three decisions, it essentially shows:

- The later date it is, the higher amount of bikes are likely to be rented
- In the late period, if there is warm weather, more bikes are likely to be rented

The error is quite big in this tree, but that is because it is *squared error*. With such a big number of bikes being rented each day, the squared error will be big. In table 2 these errors are rooted to be more interpretable. For the **node # 2**, the error is small at 660, but it increases as also total average number of rented bikes increase. To see how well the tree predicts by these simple decisions, there has been added a percentage of value, to see how big the error is per leaf node. The lowest is the last node, with a 16% error.

4 RuleFit - Task 3

4.1 Method

To solve this task in python, a RuleFit library was downloaded from RuleFit from ChristophM at GitHub. There was some problems with the debugging of visual studio code, so this code had to be in a seperate file. The dataset was preprocessed as in task 2 and in XAI1.

Node #	2	3	5	6
Rooted Squared Error	660	1014	1591	1091
Value (predicted $\hat{y}$)	1798	3934	4408	6634
Error (% of Value)	37%	26%	36%	16%

Table 2: Task 2 - Rooted squared errors, predicted values, and error percentages for each leaf node of the regression tree

Then, to specify the maximum depth to 2, an explicit *GradientBoostingRegressor* is trained as the following:

```
gb = GradientBoostingRegressor(n_estimators=500, max_depth=2, learning_rate=0.01, random_state=13)
```

To make sure that not some features gets more weight than the others, the values are scaled using *StandardScaler()* from *sklearn*. The RuleFit model is then trained with this tree as its tree generator. The rules were extracted, and the top 4 ones are found.

4.2 Results

The top four rules were found and plotted in increasing manner, as shown in figure 3.

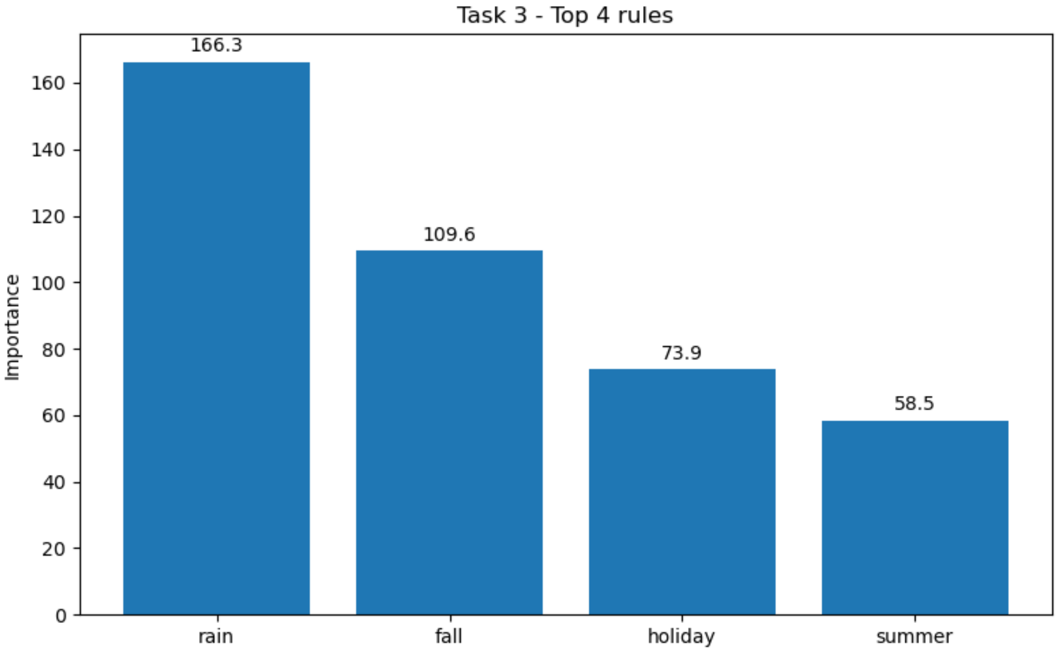

Fig. 3: Task 3 - Top 4 rules frrom RuleFit Model

4.3 Discussion

When using RuleFit, there is found that the top linear features do no longer include *days_ since_ 2011*, but rather **rain, fall, holiday** and *summer*. This differs from the task 2, but this is likely due to the Lasso regularization of the RuleFit functionality. RuleFit prefers sparse models, and it seems these four variables are combined more strong that *days_ since_ 2011* alone.

5 Conclusion

This paper has analyzed three different XAI methods, Logistic Regression, Regression Tree and RuleFit. Logistic Regression is used for binary prediction, and the results showed how different factors have different importance for the prediction of being labeled as high risk.The Regression Tree and RuleFit concluded in quite different rules being important to determine the likely amount of bikes rented, which showcased how different AI/ML methods can result in quite different results.