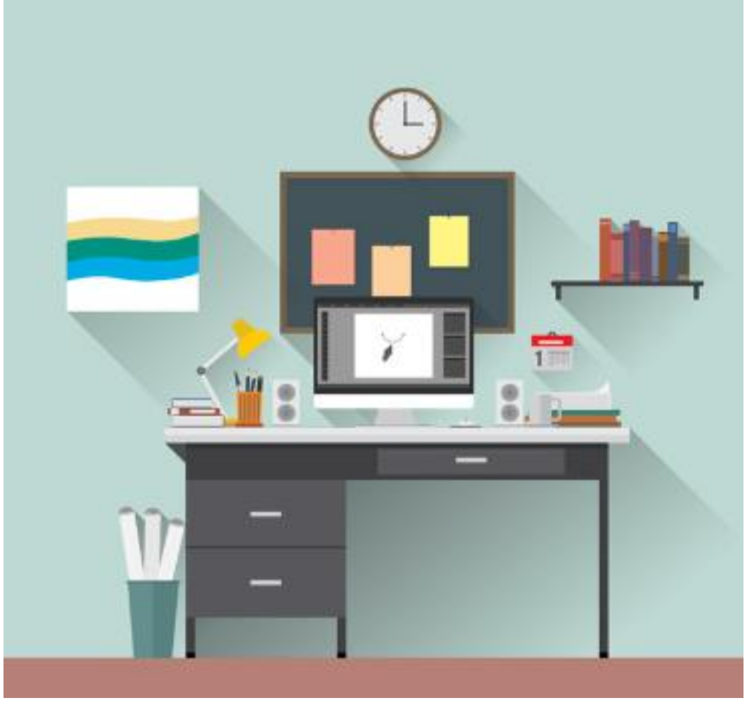
Individual Case Assignment

Machine Learning II 2018. Prof **Jesús Renero Quintero**

Global Master in Business analytics and Big Data

Absenteeism at work.analysis.



Daniel López Macías – GMBD 2018-2019

Exploratory Data Analysis.

Every data-analytics project should start with a complete and thorough analysis on the dataset object of study. It is for this reason that the dataset "absenteeism at work" has been scrutinised in a search of errors, missing values and other abnormalities that could affect our goal: to determine – based on a certain number of attributes – which workers are more likely to be absent from their work.

Running a first analysis using DSS as scrutiny method, it has been found that there is total completeness in the dataset, there are no missing values nor noise in the dataset. By running a quick modelling of the data, we can observe several factors which are the most common in the dataset. From the dataset we can infer – attending to the purposed research question: determining which individuals are more likely to be absent under certain circumstances – that this set of individuals are 36 years old and tend to become absent from work due to *"Injury, poisoning and certain other consequences of external causes "*mainly on Tuesdays and specially during the month of June. They tend to have 1 child and spend around 220 euros to get to work. This so called "Average worker" is absent from his work spot for around 7 hours annually very far away from the "most absent worker" which was away for 120 hours.

Focusing on our target variable – absenteeism time in hours – we find a very skewed data set with the majority of the values ranging between 1-10 hours and some outliers in the 120 hours zone. This tells us that the dataset does not follow a normal distribution, which could potentially limit the assumptions we infer from the dataset.

Once the dataset has been completely inspected and assured there are no errors, noise or any other deficiencies, the work can begin.

Baseline.

Once the data has been analysed it is time to get our hands on the data, first,, it is important to baseline the data, this will allow us to set realistic goals and to measure the progress obtained. As mentioned in latter lines, the absent average hours for any certain individual from the dataset is 7 hours. This will be our first starting point; it will be considered that those individuals who are away from work more than 7 hours will be considered as absent workers and those below or equal to 7 hours will not be considered as been absent, since it is inevitable that in some occasions, workers leave their work duties to attend any kind of personal matters. Running our first baseline model, the following results were obtained:

As it can be seen from the confusion matrix from Model 1 the results presented are not considered to be too disappointing. This first model was constructed using a simple linear regression model with all the independent variables presented in the dataset. Considering the requirements of the assignment it is believed that -apart from the accuracy score – the score which will be taken into consideration and trying to increase will be the precision score.

This score is considered the most significant since it assigs the precision from categorising the Absentees. As we see from the confusion matrix presented in model 1 there where a total of 47 absentees in the model (33+14) out of which only 14 where misclassified. This would mean that the model correctly determined 34 absentees and classified 13 absentees as non-absent which where absentees. Similarly, the model classified 20 non-absentees as absentees. It is considered this is not a big problem since if we extrapolated this problem to real life and determined that on a certain day, X numbers of workers were going to show up to work and more of them appeared, this would not pose a problem for the company, however, if we expected X workers to show and fewer appeared we would have problems.

Model 1

Seed	3457
Threshold	0.5
Training Dataset	80%

		Reference	
		0	1
Prediction	0	81	14
	1	20	33

Accuracy	0.77
Precision	0.70
Recall	0.62
Specificity	0.85

Feature Engineering.

Based on the insights obtained from the elaboration of model 1, some model engineering was carried out to enhance the results obtained. The first modification carried out was the transformation of all the variables into integers, to make sure there was a consistency in the data frame. This included the transformation of the variable "Work load Average day" which was stored as a factor and therefore many levels were present in the data set, distorting the reality of the data. The second step was to normalise the variables, such as "weight" or "age" did not have a higher weight within the data compared to those dichotomous variables. Lastly, it was decided to remove the variables "weight" and "height" since there was a high correlation of both with the variable "body mass index". Pearson's test was conducted resulting in the rejection of the null hypothesis with a 95% confidence. Based on the modifications mentioned latterly Model 2 (please refer to Evaluation and Validation was composed. Assessing the new model performance, it is observed how metrics have dropped around a 5%. Nevertheless, recall values have remained unchanged, therefore the model is misclassifying the same number of individuals at a lower computational expense. To keep enhancing the model a backward model selection was run to select those variables that achieved the lowest AIC value -1348 was the value obtained for the variables described in Appendix 1 Based on the work abovementioned, it is believed that the data set - resulting from the previously described transformations – contains the variables that will allow the best classification results.

Evaluation and Validation

Taking the data frame created from the transformations described in the Feature engineering section, the following methods will be tested to determine which one offers the best classification model. The models used were:

- Logistic regression model
- Lasso
- Ridge
- KNN

To make sure the models were comparable all of them were analysed using a fixed seed value of 3457. The results obtained are presented below:

Linear Regression Model

Seed	3457
Threshold	0.5
Training Dataset	80%

		Reference	
		0	1
Prediction	0	97	24
	1	30	49

Accuracy	0.73
Precision	0.67
Recall	0.62
Specificity	0.80

Ridge Model

Seed	3457
Threshold	0.5
Training Dataset	80%

		Reference	
		0	1
Prediction	0	97	24
	1	35	44

Accuracy	0.71
Precision	0.65
Recall	0.56
Specificity	0.80

Lasso Model

Seed	3457
Threshold	0.5
Training Dataset	80%

		Reference	
		0	1
Prediction	0	97	24
	1	33	46

Accuracy	0.72
Precision	0.66
Recall	0.58
Specificity	0.80

KNN Model

Seed	3457
Threshold	0.5
Training Dataset	80%

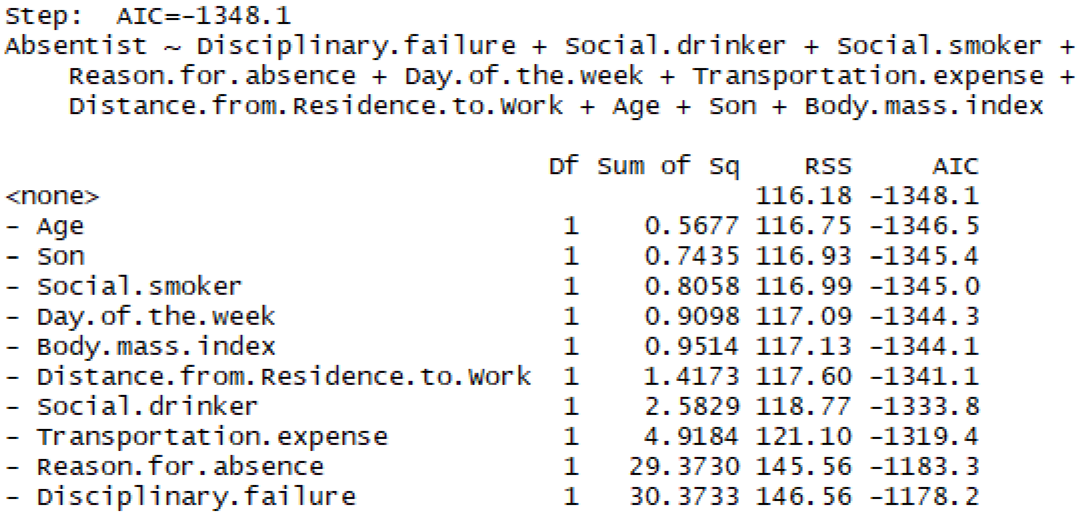
		Reference	
		0	1
Prediction	0	117	4
	1	33	46

Accuracy	0.82
Precision	0.92
Recall	0.58
Specificity	0.97

Based on the data presented, it is clear that the best classification method is using KNN, since it provides a precision of 82% and a recall of 92%.The model was trained using the variables presented in Appendix 1 as so were the rest of the models.

Appendix.

1.



```
Step:  AIC=-1348.1
Absentist ~ Disciplinary.failure + Social.drinker + Social.smoker + 
    Reason.for.absence + Day.of.the.week + Transportation.expense + 
    Distance.from.Residence.to.Work + Age + Son + Body.mass.index

                                  Df Sum of Sq    RSS     AIC
<none>                                         116.18 -1348.1
- Age                              1    0.5677 116.75 -1346.5
- Son                              1    0.7435 116.93 -1345.4
- Social.smoker                    1    0.8058 116.99 -1345.0
- Day.of.the.week                  1    0.9098 117.09 -1344.3
- Body.mass.index                  1    0.9514 117.13 -1344.1
- Distance.from.Residence.to.Work  1    1.4173 117.60 -1341.1
- Social.drinker                   1    2.5829 118.77 -1333.8
- Transportation.expense           1    4.9184 121.10 -1319.4
- Reason.for.absence               1   29.3730 145.56 -1183.3
- Disciplinary.failure             1   30.3733 146.56 -1178.2
```