berlin international
UNIVERSITY OF APPLIED SCIENCES

Project Report

Analytics Lab 2: Machine Learning 23/24

Jaeden Capinig

2202987

Introduction

Businesses are all about efficiency and saving costs. Even with the rise of AI, we are still dependent on human capital for running business operations. Unlike a machine, humans are prone to accidents, sickness and circumstances out of their control, leading them to be absent in the workplace. Hours not worked equates to loss of productivity and opportunity to generate profits. If only companies could predict the number of hours workers will be absent, they could probably estimate loss in advance, and find ways to compensate for it.

We would explore if certain features could predict the number of hours a worker would be absent.

Relevant libraries imported are:

- pandas
- numpy
- seaborn
- matplotlib.pyplot
- statsmodels.formula.api
- statsmodels.api
- sklearn.linear_model
- math
- matplotlib

Understanding the data

My chosen dataset is publicly available in the UC Irvine Machine Learning Repository (archive.ics.uci.edu).
Records of absenteeism were sourced from a courier company in Brazil, within a time period of July 2007 to July 2010 (3 years). It was used in academic research at the Universidade Nove de Julho, Postgraduate Program in Informatics and Knowledge Management.

In the data set, there are a total of 21 variables and is split by the following:

Categorical variables

- Individual identification (ID)
- Reason for absence (ICD) - Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI). See Appendix A for comprehensive details

- Month of absence
- Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
- Seasons (summer (1), autumn (2), winter (3), spring (4))
- Disciplinary failure (yes = 1; no = 0)
- Social drinker (yes = 1; no = 0)
- Social smoker (yes = 1; no = 0)

Ordinal Variable

- Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

Numerical Variables

- Transportation expense
- Distance from Residence to Work (kilometers)
- Service time
- Age
- Work load Average/day
- Hit target
- Son (number of children)
- Pet (number of pets owned)
- Weight
- Height
- Body mass index

Dependent or target variable

- Absenteeism time in hours

The following is observed of the 740 entries:

- There are 0 values for the columns ‘Reason for absence’, ‘Month of absence’, and ‘Absenteeism time in hours’. Assuming there are workers that were never absent, we will think about if we will remove them from the dataset
- Average day workers are absent is on a Wednesday (3.91 or 4)
- Workers travel an average distance of 29.63 km from home to work
- Mean age is 36.45
- average absenteeism time is 6.92 hours.

Preparing the dataset

To prevent any problems in running the code (i.e. spaces found in column names), we changed the name of the following variables:

- 'Reason for absence' to 'reason_absence',
- 'Month of absence' to 'month',
- 'Day of the week' to 'day_of_week',
- 'Seasons' to 'season',
- 'Transportation expense' to 'transpo_exp',
- 'Distance from Residence to Work' to 'home_work_km',
- 'Service time' to 'work_hours',
- 'Work load Average/day ' to 'work_avg',
- 'Disciplinary failure' to 'discipline',
- 'Education' to 'education',
- 'Son' to 'no_children',
- 'Social drinker' to 'social_drinker',
- 'Social smoker' to 'social_smoker',
- 'Body mass index' to 'bmi',
- 'Absenteeism time in hours' to 'hours_absent'

Visualizing values within categorical variables

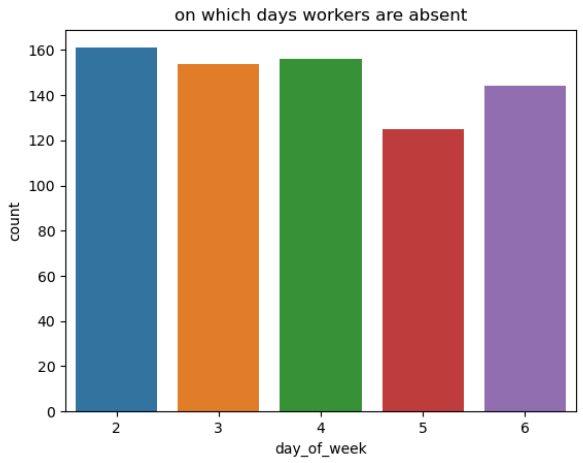


figure 1.1: 2 (Mon.), 3 (Tues.), 4 (Wed.), 5 (Thurs.), 6 (Fri.)

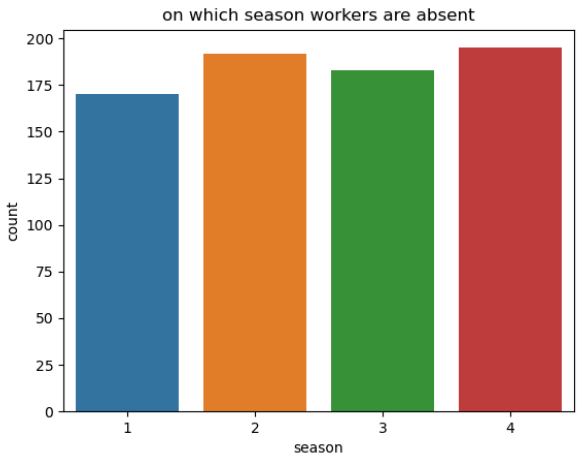
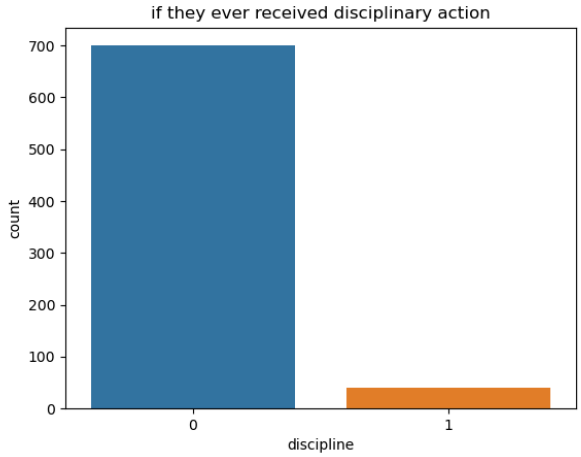


figure 1.2: 1 (Summer), 2 (Autumn), 3 (Winter), 4 (Spring)



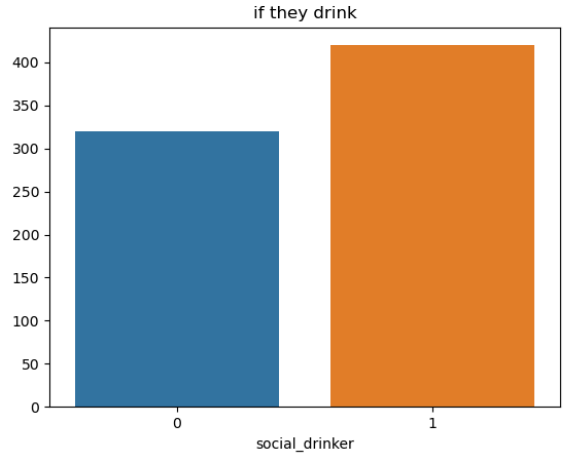
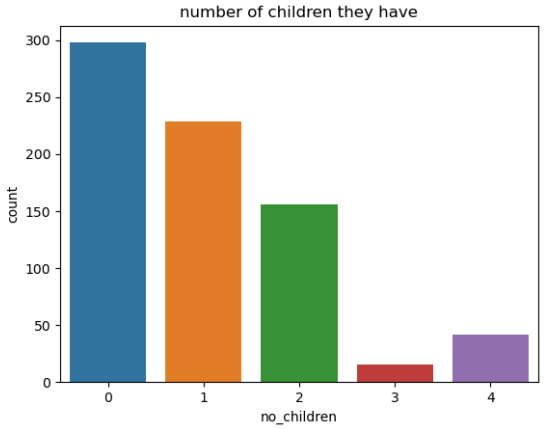


figure 1.3 & 4: 0 (No), 1 (Yes)



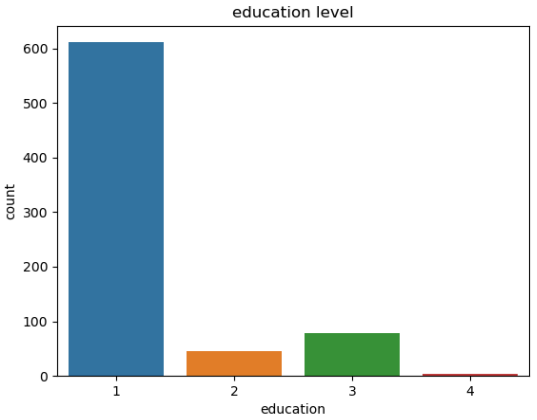


figure 1.6: 1 (high school), 2 (graduate), 3(postgraduate), 4 (master and doctor)

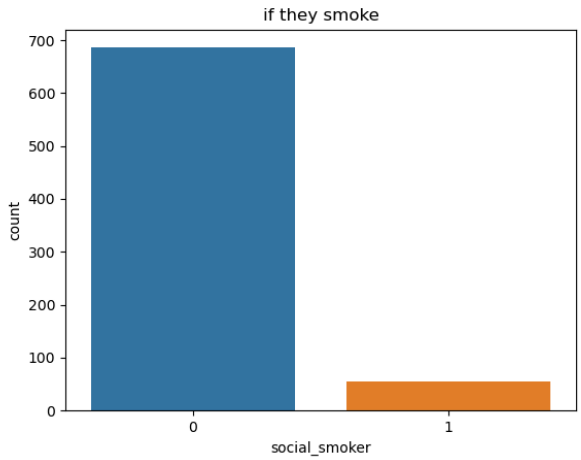


figure 1.7: 0 (No), 1 (Yes)

Selecting the interesting data above:

- 82.87% of the workers are high school graduates
- 40.27% of them have no children
- only 7.3% of them smokes

There are 43 entries where **`reason_absence`** has the value of 0, 3 entries for **`month`**, and 44 entries for **`hours_absent`**. We decided to keep all of them, as they may still be valuable in training and testing

Bivariate analysis

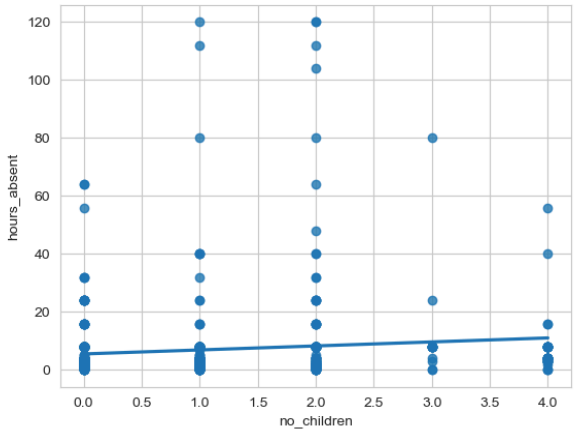


figure 2.1

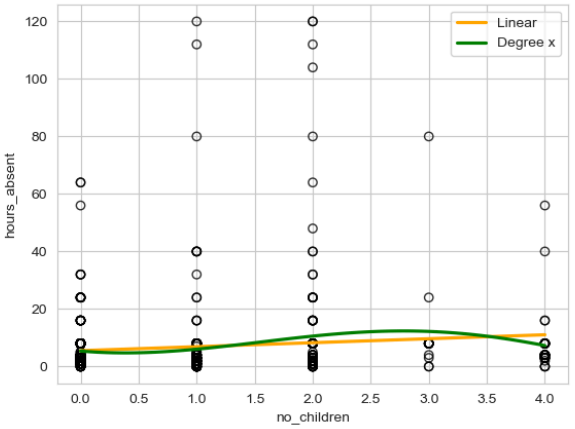


figure 2.2

In the linear regression plot, the linear regression line is relatively flat, indicating a weak linear relationship between the number of children a worker has and the number of hours they are absent. Spread of data points suggests variability, but lacking in either strong upward or downward trend. The Polynomial regression plot also suggests a weak relationship between **`no_children`** and **`hours_absent`**.

Running the Ordinary Least Squares, we had the following results:

- R-squared of 0.013 implies that no_children explains only 1.3% of the variability in hours_absent.
- P-value is statistically significant (0.002)
- Positive coefficient of 1.38 suggests that the more children a worker has, the more hours absent he'll be.

- High skewness and kurtosis suggests non-normality of the residuals.
- Therefore, model fit might not be good

Looking at the correlation matrix, the next noteworthy variable is **`transpo_exp,`** which has a positive correlation of 0.028, followed by **`work_avg`** with 0.025.

Checking out the standardized regression coefficients, we had a result of **`[-0.0037036 0.00733949 1.45971697].`** Values indicate that as the number of children increase by one standard deviation (1.1), hours absent increase by 1.46 standard deviations. Standard deviation of **`hours_absent`** is 13.33 so constituting a change of 19.46 hours(13.33 x 1.46).

In running the Generalized Linear Model Regression Results, we received the following results:

- Standard errors and coefficients are extremely large (e.g. intercept is **`8.679e+16`**)
- Z- and p-values are 0, but don't reflect true significance. rather, serious model fit issues.

Resampling

To remedy found issues in my dataset, bootstrapping could help in assessing stability and variability of the model coefficients, while cross-validation evaluates model performance and generalizability.

2-fold cross validation

We have a mean squared error or MSE of 192.93, the highest among the 3 approaches. This projects a poor model fit.

Leave one out cross validation

Even though LOOCV provides a nearly unbiased estimate model performance, MSE is 176.09, and standard deviation of the errors is 1114.97, which is still relatively high.

K - Fold cross validation

Across 10 polynomial degrees, MSE was varying around 176, and standard deviations ranging from 79 to 83. This did not improve model performance

Bootstrapping

conventional linear regression yields an intercept of 5.52, with coefficient of 1.38

Bootstrapping has an intercept of 4.54, with coefficient of 2.05

High MSE values from different cross-validation methods suggest that predictor **`no_children`** might not directly influence target variable **`hours_absent.`**

As there is a difference in estimates between a conventional linear model regression and bootstrapping, the model might not be stable and could have high variability. Although running the code multiple times gets it closer to the estimates in the lm model.

Simple Linear Model for prediction (Lasso)

We will only perform Lasso, as there are only a few possibly good predictors found in our dataset (i.e. **`no_children, work_avg, transpo_exp`**), and too many included.

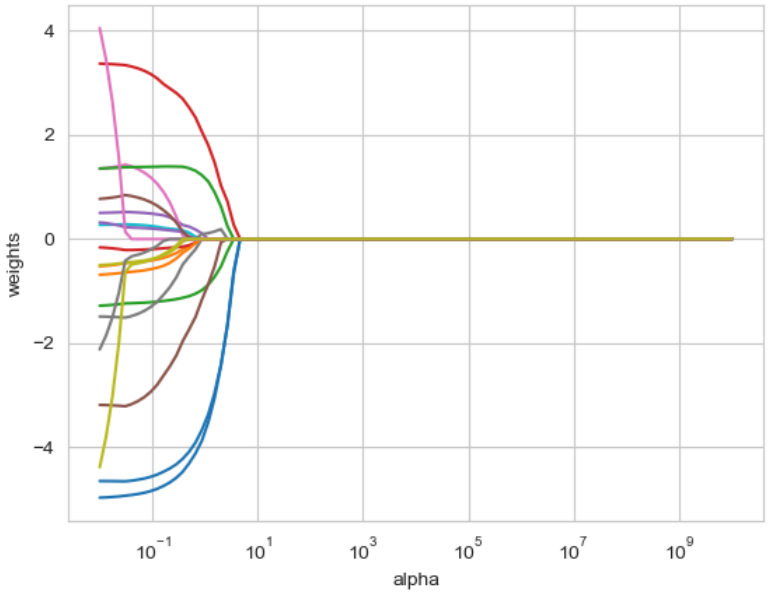


figure 3: as alpha increase, more coefficients shrink to zero

In printing the coefficients, most significant predictors are the following:

- **`social_drinker`**: 1.94
- **`no_children`**: 1.18

- `transpo_exp`: 0.059
- `home_work_km`: -1.03

Non-linear Models

Random Forests

I started with this method as this outperforms three other methods in terms of having a balance between performance and interpretability.

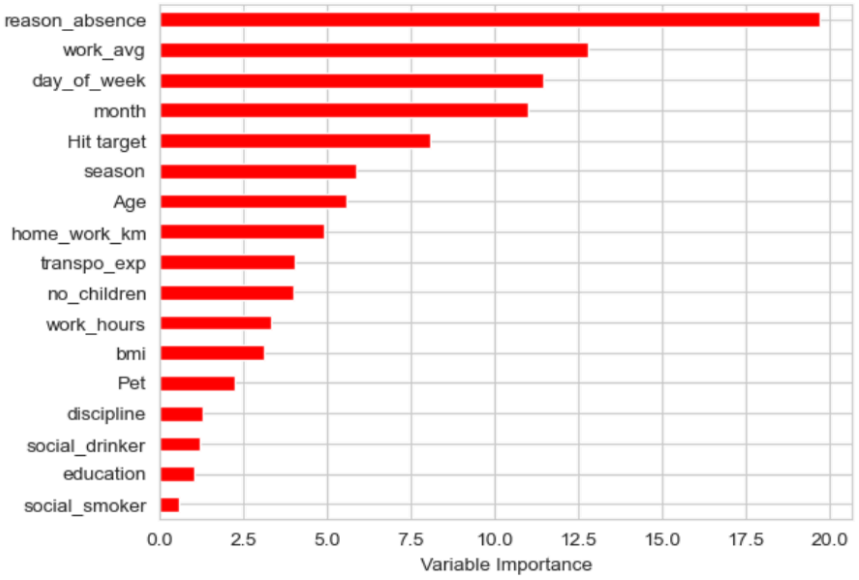


figure 4.1

As shown above, the most significant predictor is reason_absence (<19), which we overlooked previously. The next is work_avg (> 12.5), and no_children (> 4) is way below the hierarchy which we keep considering.

Boosting

This method offers the highest accuracy, but requires careful tuning.

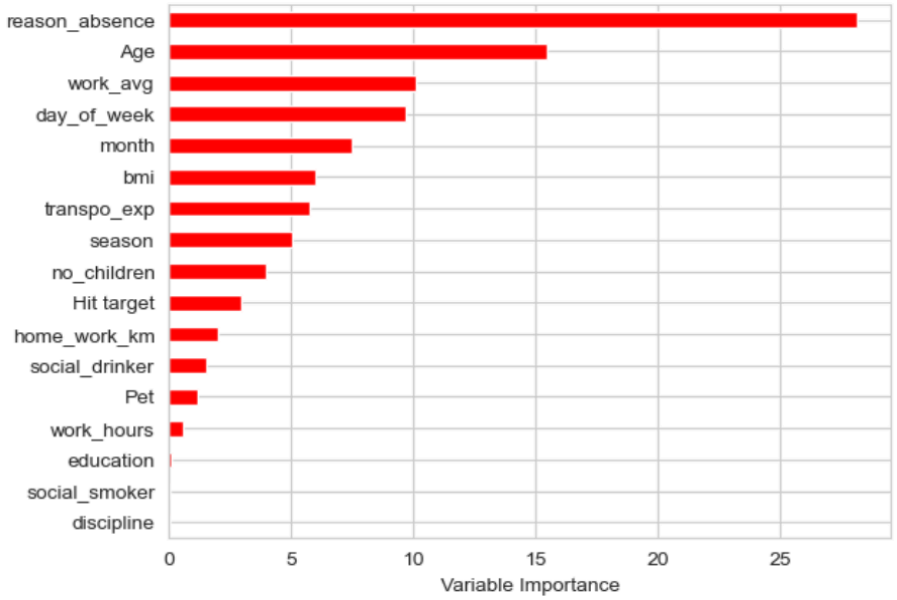


figure 4.2

In the feature importance plot, most important are **`reason_absence`**, **`Age`**, and **`work_avg`**

In model performance, when learning_rate is at 0.01, our MSE is 114.26. Further iterations are encouraged to determine optimal performance.

Overall, there could be underlying problems with the dataset, even though we saw better results in running Resampling and non-linear models.

We also overlooked variable **`reason_absence`** as a potential good predictor for **`hours_absent.`**

Reference

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). *Application of a neuro fuzzy network in prediction of absenteeism at work*. 1–4.

Appendix A

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases
II Neoplasms
III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV Endocrine, nutritional and metabolic diseases
V Mental and behavioral disorders
VI Diseases of the nervous system
VII Diseases of the eye and adnexa
VIII Diseases of the ear and mastoid process
IX Diseases of the circulatory system
X Diseases of the respiratory system
XI Diseases of the digestive system
XII Diseases of the skin and subcutaneous tissue
XIII Diseases of the musculoskeletal system and connective tissue
XIV Diseases of the genitourinary system
XV Pregnancy, childbirth and the puerperium
XVI Certain conditions originating in the perinatal period
XVII Congenital malformations, deformations and chromosomal abnormalities
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX Injury, poisoning and certain other consequences of external causes
XX External causes of morbidity and mortality
XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).