NEXT STEPS

1. Prepare the ML algorithms:
   1. Select and compare multiple algorithms
   2. Complete a 10-fold Cross-validation on sets and select the highest performing model. Explain your decision.
2. Train and tune the chosen model.
3. Present the solution, including a detailed report of findings similar to the first three assignments.
4. Launch the model

You will use simple *k-fold cross-validation* to evaluate the performance of any models you choose. Use any models depending on your research question (k-means clustering, support vector, random forest, etc.)  and select the highest performing model to complete your analysis. Explain your model evaluation and why you chose one over the other; complete a model evaluation, use some metric to select the model (RMSE or others).

**Introduction**

Absenteeism refers to an employee’s habitual absence from their place of work. There are many factors that impact why an employee may miss days of work at a time. For example, they could have a medical condition, live far from work, etc. In predicting absenteeism, companies can maximize profit by hiring a workforce that shows up to work often. However, considering factors such as age, sex, number of children, level of education, and other personal data could lead to bias and unfair hiring practices. Therefore, in my analysis of the data set of 740 employees from a Courier company, I will develop a model to predict absenteeism based on only features that are descriptive of the absence itself, not the employee or any other aspect of their social lives.

**Exploratory Data Analysis**

To begin my analysis, I performed a random forest test on all of the features in the dataset to ensure that my intuition was not ignoring any obvious predictors in the dataset. The results indicated that the most important features for predicting absenteeism in the dataset are:

1 Reason for absence int64

2 Month of absence int64

3 Day of the week int64

4 Seasons int64

5 Transportation expenses int64

9 Workload float64

10 Hit target int64

My intuition was correct in the sense that none of the personal or social features are of utmost importance in predicting absenteeism in this data.

To dive deeper into the analysis, I created a correlation matrix to understand how all of the non-personal (other than the reason for absence, which I consider being a description of the absence itself, although it may reveal personal information) and non-social features. For these purposes, the following columns are excluded from the graph below: 'Pet','Weight', 'Age', 'Social drinker', 'Social smoker', 'Height', 'Body mass index', 'Son', 'ID'.

**A picture containing chart

Description automatically generated**

In my exploratory analysis, I found that the remaining features have interesting relationships. ‘Month of absence’ and ‘Seasons’ are positively correlated, which makes sense because the seasons change as the month changes. For my analysis, I will only include Month of absence, which can be a proxy for seasons. I chose the Month over the seasons because it is more granular and easily interpretable. Similarly, ‘distance from work’ and ‘transportation expense’ are related and arguably proxies of each other. I will likely choose distance from residence to work over transportation expense because I am not sure if the expense variable is monthly or daily expense, and ‘distance from residence to work’ has a stronger correlation with ‘reason for absence’.

I then ran another random forest on only the features that I am interested in to determine which is the stronger/more important predictor. I included ‘disciplinary failure’ because it had a negative correlation with reason for absence in the correlation matrix, but according to the random forest output, it doesn’t seem to be that important.

1 Reason for Absence: Score: 0.24653

2 Month of Absence: Score: 0.10845

3 Day of the Week: Score: 0.13435

4 Distance from Residence to Work: Score: 0.12049

5 Service time: Score: 0.12133

6 Workload: Score: 0.12781

7 Hit Target: Score: 0.09776

8 Disciplinary Failure Score: 0.04328

**Descriptive Statistics for Features of Interest**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | Reason for absence | Month of absence | Day of the week | Distance from Residence to Work | Service time | Hit target | Disciplinary failure | Absenteeism time in hours | Workload |
| **count** | 740 | 740 | 740 | 740 | 740 | 740 | 740 | 740 | 740 |
| **mean** | 19.21621622 | 6.324324324 | 3.914864865 | 29.63108108 | 12.55405405 | 94.58783784 | 0.05405405405 | 6.924324324 | 271490.2351 |
| **std** | 8.433405883 | 3.436286932 | 1.42167471 | 14.83678844 | 4.384873408 | 3.779313134 | 0.2262772732 | 13.3309981 | 39058.11619 |
| **min** | 0 | 0 | 2 | 5 | 1 | 81 | 0 | 0 | 205917 |
| **25%** | 13 | 3 | 3 | 16 | 9 | 93 | 0 | 2 | 244387 |
| **50%** | 23 | 6 | 4 | 26 | 13 | 95 | 0 | 3 | 264249 |
| **75%** | 26 | 9 | 5 | 50 | 16 | 97 | 0 | 8 | 294217 |
| **max** | 28 | 12 | 6 | 52 | 29 | 100 | 1 | 120 | 378884 |

**Non-normal Outcome variable**

‘Absenteeism in hours’ is not a normally distributed variable, but this is okay because regression does not require normality. However, a standard scaler will be used to scale all data when I am testing different models in my research. I am choosing analysis regression because of the nature of my outcome variable and my descriptive features are mostly categorical.

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

A picture containing chart

Description automatically generated

**Conclusion**

Predicting absenteeism is important for any company but including certain data about employees could lead to discriminatory hiring practices. My research questions are supported by the data, as the Random Forest analysis indicated that the most important features for predicting absenteeism are what I categorize as non-personal, non-social features (i.e. social drinker, number of children, age, etc.) My analysis and work toward a final model will investigate how to most accurately predict absenteeism using non-personal and non-social features. I will compare my models to models that include more pervasive features like age, number of children, social drinking, etc.

**Appendix A: Attribute Information:**

1. Individual identification (ID)
2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) are 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).

1. Month of absence
2. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
3. Seasons
4. Transportation expense
5. Distance from Residence to Work (kilometers)
6. Service time
7. Age
8. Workload Average/day
9. Hit target
10. Disciplinary failure (yes=1; no=0)
11. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
12. Son (number of children)
13. Social drinker (yes=1; no=0)
14. Social smoker (yes=1; no=0)
15. Pet (number of pets)
16. Weight
17. Height
18. Body mass index
19. Absenteeism time in hours (target)

.arff header for Weka:

@relation Absenteeism\_at\_work

@attribute ID {31.0, 27.0, 19.0, 30.0, 7.0, 20.0, 24.0, 32.0, 3.0, 33.0, 26.0, 29.0, 18.0, 25.0, 17.0, 14.0, 16.0, 23.0, 2.0, 21.0, 36.0, 15.0, 22.0, 5.0, 12.0, 9.0, 6.0, 34.0, 10.0, 28.0, 13.0, 11.0, 1.0, 4.0, 8.0, 35.0}

@attribute Reason\_for\_absence {17.0, 3.0, 15.0, 4.0, 21.0, 2.0, 9.0, 24.0, 18.0, 1.0, 12.0, 5.0, 16.0, 7.0, 27.0, 25.0, 8.0, 10.0, 26.0, 19.0, 28.0, 6.0, 23.0, 22.0, 13.0, 14.0, 11.0, 0.0}

@attribute Month\_of\_absence REAL

@attribute Day\_of\_the\_week {5.0, 2.0, 3.0, 4.0, 6.0}

@attribute Seasons {4.0, 1.0, 2.0, 3.0}

@attribute Transportation\_expense REAL

@attribute Distance\_from\_Residence\_to\_Work REAL

@attribute Service\_time INTEGER

@attribute Age INTEGER

@attribute Work\_load\_Average/day\_ REAL

@attribute Hit\_target REAL

@attribute Disciplinary\_failure {1.0, 0.0}

@attribute Education REAL

@attribute Son REAL

@attribute Drinker {1.0, 0.0}

@attribute Smoker {1.0, 0.0}

@attribute Pet REAL

@attribute Weight REAL

@attribute Height REAL

@attribute Body\_mass\_index REAL

@attribute Absenteeism\_time\_in\_hours REAL