

|  |  |
| --- | --- |
| Term Project BEHAVIOURAL ANALYSIS OF ABSENTISM AT WORK | |
|  |  |
| CSIS-3290-001 FUNDAMENTAL OF MACHINE LEARNING JULY 06, 2020HARSHIT BANSAL 300303010 |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | |  |
| DISCOVERY **PROBLEM** | | | |
| ABSTRACT The database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.  The dataset contain attribute that explain the different reasons for not making up on the job.  I choose this database so as to get insights and also to share this insight with any organization that I might work with. These analytics will provide me a satisfactory goal of factors responsible for absenteeism. | | [This Photo](https://commons.wikimedia.org/wiki/File:Wikimood_absent.png) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)  To find out factors affecting absenteeism at work. | |
|  | |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | TO STUDY THE FACTORS AFFECTING ABSENTEESM WE ASSUME THAT DISTANCE FROM RESIDENCE PLAY AN IMPORTANT ROLE IN THE STUDY MENTIONED ABOVE. ALSO, WE ASSUME THAT THERE IS A HIKE IN ABSENTEESM DURING HOLIDAY SEASON. WE WILL ANALYSE AND SEE IF OUR HYPOTHESIS IS CORRECT. HYPOTHESIS |  |
|  | POTENTIAL DATA SOURCE  https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work# |  |

|  |  |
| --- | --- |
|  | There are 21 columns from which I have to find what column affect the Number of Absent hours by studying the available data.  **Data Preparation**   1. ID 2. Reason of Absence 3. Month of absence 4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)) 5. Seasons 6. Transportation expense 7. Distance from Residence to Work (kilometers) 8. Service time 9. Age 10. Work load Average/day 11. Hit target 12. Disciplinary failure (yes=1; no=0) 13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4)) 14. Son (number of children) 15. Social drinker (yes=1; no=0) 16. Social smoker (yes=1; no=0) 17. Pet (number of pet) 18. Weight 19. Height 20. Body mass index 21. Absenteeism time in hours (target)   The goal of my project is pretty simple hence you will not see complex code or different models with diagrams because as I mentioned above my project is it see what columns or factors affect the number of absent hours. This data is from s courier company in brazil and is collected few years back. Since it is not that recent things might have changed but I had no option, this is the only available data. Further I will explain my key findings and how I get them. Please don’t expect complex code. |
|  | **Data Processing** |

In this part of the document I will explain the data processing for my term project Absenteeism data that I selected. Since I the categorical data is pre-converted into integer data in my dataset, I don’t have to do much data processing as generally I found out from my research that a data scientist spends 60% of the project resources in cleaning the data.

Since Day of the week is converted into numeric form It was okay for me to not convert it into dummy variable as data correlates the different rows. That is my assumption. Also, we have studied similar kind of data during the coursework.

I dropped ID column because it was irrelevant to our analysis.

**Model Planning and Implementation**

I propose linear regression model as I study the data, it’s not categorical also I see linear regression as the best model to use for Absenteeism at work data.

Though I tried Linear and SVM model for my data to predict results but unfortunately the accuracy score for both the data was same hence I was in a bit dilemma which model to use. I could have tried other models as well but during my research I found out that the rest of the models are not that fit for analysis in light to Absenteeism at work data that I am using for my analysis.

The end goal for this project is to predict the factors affecting number of absent hours for employee and to analyze these factors for future employee to be hired for the industry.

**Results Interpretation and Business Implications**

While interpretation of the results, I can see that the accuracy score for these two models are same but the results I have from the two model are quite different hence I will put down a table below to see what results we got from the two statistical models to see which one fits the best.

|  |  |  |
| --- | --- | --- |
| Models | Linear Regression | SVM |
| Reason of Absence | True | True |
| Month of Absence | True | False |
| Day of the week | True | True |
| Seasons | False | False |
| Transportation expense | False | False |
| Distance from Residence to Work | False | False |
| Service time | True | False |
| Age | False | False |
| Workload Average/day | True | True |
| Hit target | True | False |
| Disciplinary failure | True | True |
| Education | False | True |
| Son | False | True |
| Social drinker | False | True |
| Social smoker | False | False |
| Pet' | False | False |
| Weight | False | False |
| Height | False | False |
| Body mass index | False | True |
| Absenteeism time in hours | False | False |

In the table above, you might see some dissimilar result from the two models but I am quite sure from the accuracy of these models and using my own logic that four factors that are common in both the model states that these four factors extremely influence on the number of absent hours of an employee. Reason of absence, day of the week, Workload average/day and disciplinary failure are the four major factors that affect the number of absent hours.

Also not ignoring the other factors that we found that may be significant are month of absence, service time, hit target, social drinker, son and BMI. From these results it makes pure sense of the factors that might affect the target variable.

**CONCLUSION**

From the analysis above we can say that there are four factors that might affect the absenteeism at work and these factors does not depend on what ethnicity you belong, how far you live or what vehicle you drive. It totally depends upon the behavior of the individual that has been employed and also how dedicated the person is to the job, no matter if he lives 20 miles away, If that person is disciplined to fulfil the duty, despite of the age, distance or whatsoever they will not get absent on the day of the work. Yes, there are few factors that might affect a little, but the core finding says what I stated above.