Capstone Project - Workplace Absenteeism

Presentation by: Sara

Workplace Absenteeism – Introduction

Absenteeism is the practice of regularly being absent from work. While employers expect employees to miss some days of work, excessive absenteeism can hurt productivity and the company's bottom line. If employers are more aware of the causes of absenteeism, they are likelier to take steps to alleviate the problems. The goal of this project is to determine the most common causes of absenteeism.

Workplace Absenteeism – Outline of Approach

- Obtain the data from the Machine Learning Repository of the University of California, Irvine.
- Wrangle the data
- Statistical and visual analysis
- Regression and predictions
- Conclusion

Workplace Absenteeism – Obtain the Data

This dataset was accessed from the from the Machine Learning Repository of the University of California, Irvine. (http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work) These are the factors of interest:

- Reason for absence - absences recognized by the International Code of Diseases (21) and not recognized by ICD (7).
- Day of the week
- Season
- Month of absence
- Transportation expense in reais (R$)
- Distance from Residence to Work (kilometers)
- Service time in years

Workplace Absenteeism – Obtain the Data (continued)

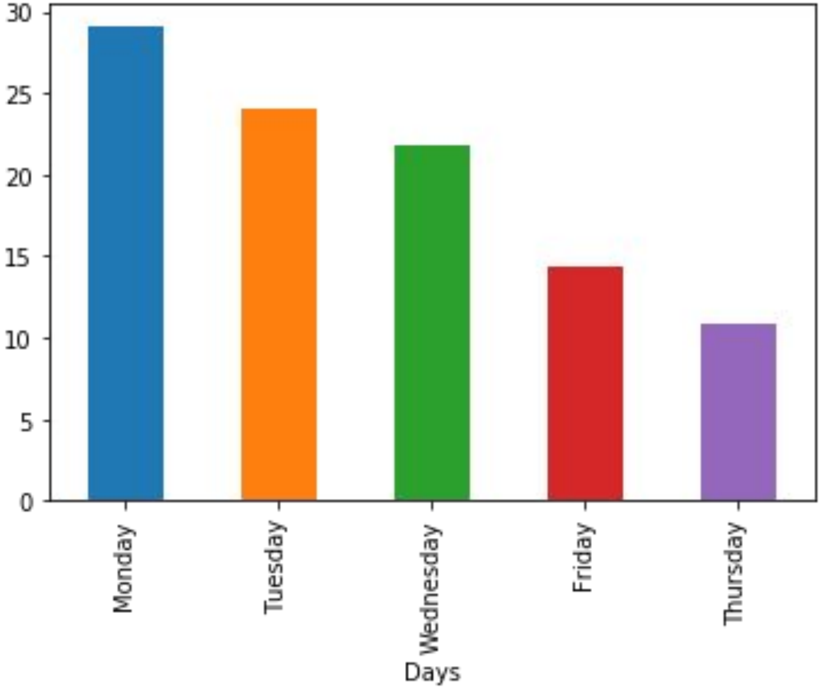
- Age (years)
- Work load Average/day
- Hit target
- Disciplinary failure (binary with 1 for yes and 0 for no)
- Education level
- Number of children
- Social drinker (binary)
- Social smoker (binary)
- Number of pets
- Weight (kg)
- Height (cm)
- Body mass index
- Number of hours absent

Workplace Absenteeism – Data Wrangling

The data is clean, so I did the following:

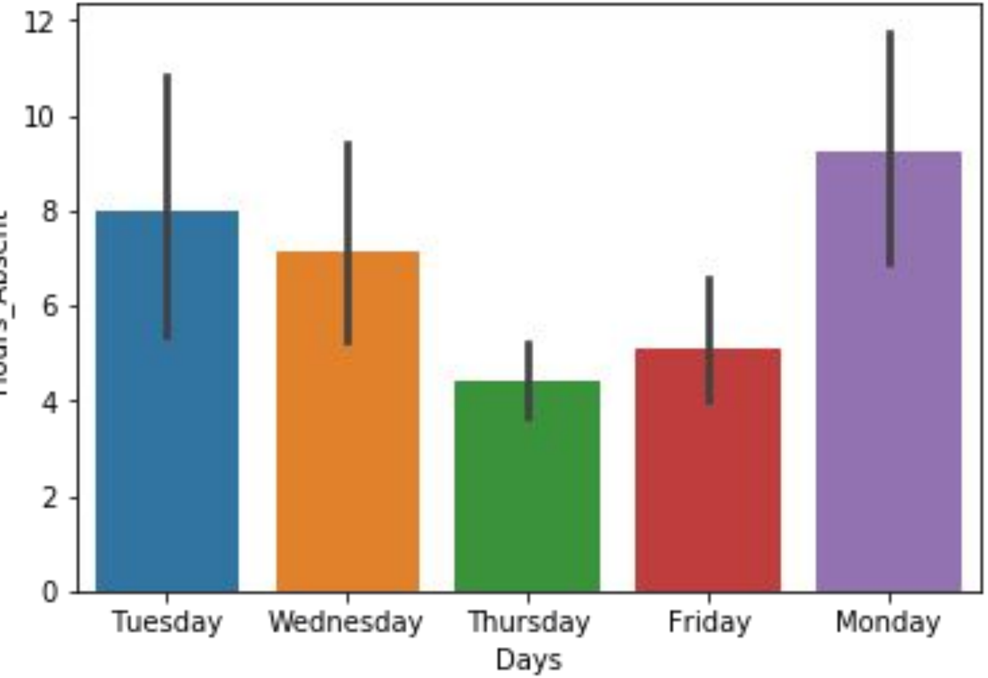
- Binary columns for regression analysis using one-hot encoding for 'Reason', 'Month of absence', 'Day of the week','Seasons', 'Disciplinary failure', 'Education', 'Social drinker', and 'Social smoker'.
- Renamed the new columns from the one-hot encoding and made the existing columns easier to understand.
- Added a new column with the hours absent logarithmically transformed and replaced the '-inf' values with zeros.

Farmer's Markets – Statistical and Visual Analysis



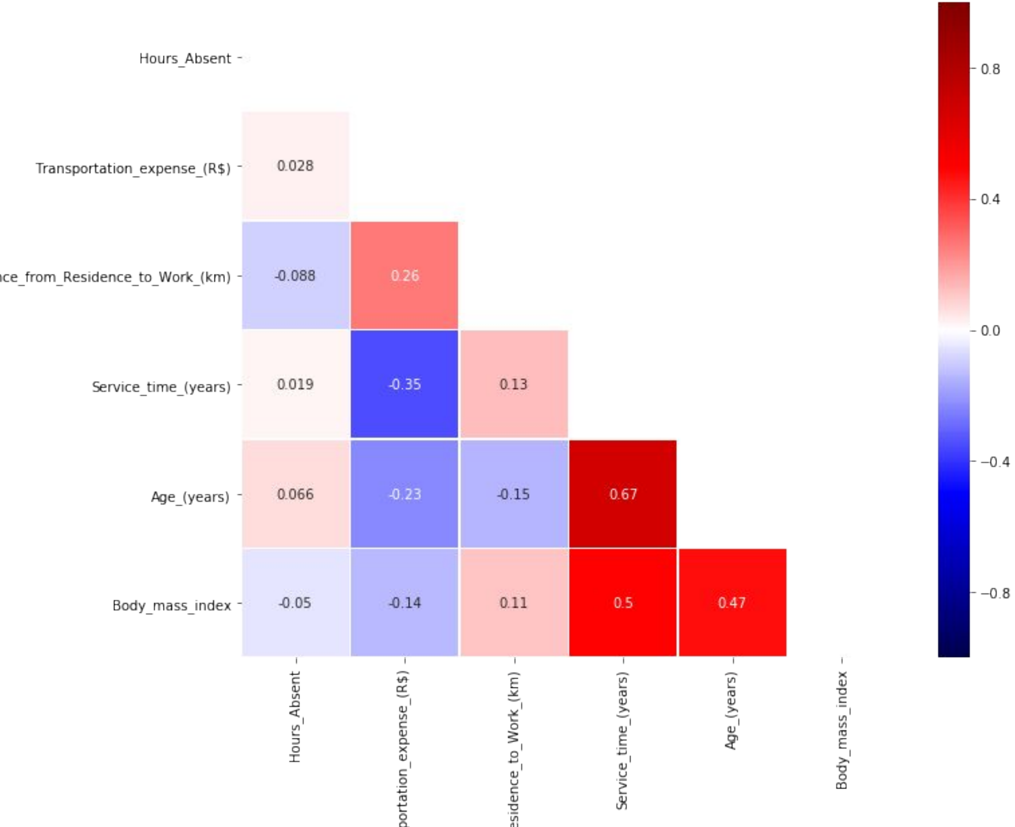
For part of the statistical and visual analysis, the percentages of hours missed for factors such as days of the week, months, seasons, and medical reasons were calculated and plotted. Here is the plot of the percentage of hours missed for each day of the week. Monday has the highest percentage of hours missed while Thursday has the least.

Farmer's Markets – Statistical and Visual Analysis (continued)



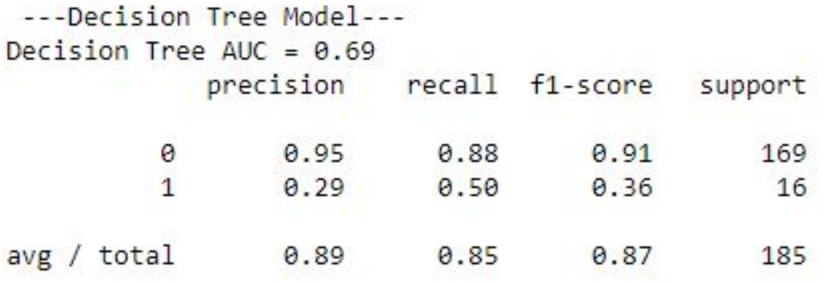
The means and standard deviations of hours missed for each variable were also calculated. Here are the calculated means and standard deviations (the black lines) of each day of the week. There is significant overlap between the days, so there is not a significant difference between the means of each day. If only Monday and Thursday were compared, then it would be clear that there was a significant difference between the two days.

Farmer's Markets – Statistical and Visual Analysis (continued)



This is the correlation matrix of each continuous variable (transit costs, distance, service time, age, BMI, and hours absent). The correlations of hours absent with the other variables are not very strong.

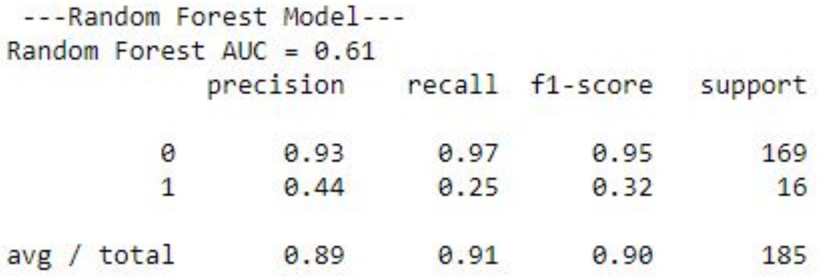
Workplace Absenteeism – Regression and Predictions



```
 ---Decision Tree Model---
Decision Tree AUC = 0.69
             precision    recall  f1-score   support

          0       0.95      0.88      0.91       169
          1       0.29      0.50      0.36        16

avg / total       0.89      0.85      0.87       185
```

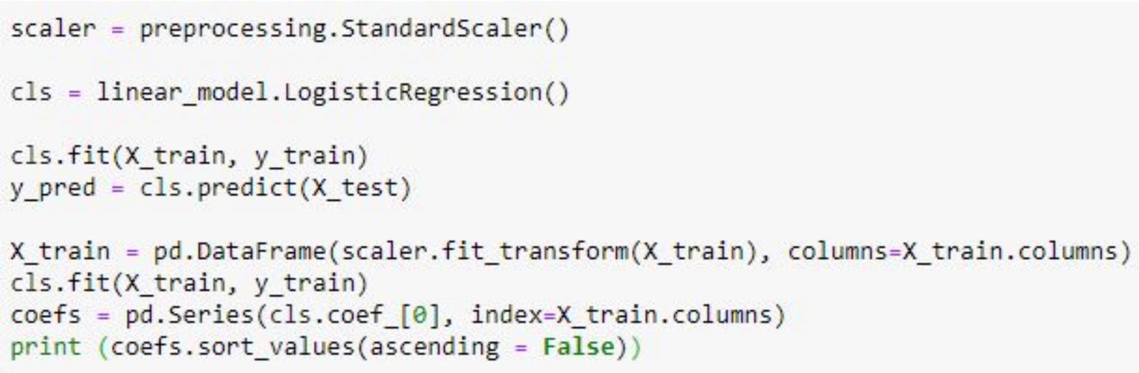


```
 ---Random Forest Model---
Random Forest AUC = 0.61
             precision    recall  f1-score   support

          0       0.93      0.97      0.95       169
          1       0.44      0.25      0.32        16

avg / total       0.89      0.91      0.90       185
```

Workplace Absenteeism – Regression and Predictions (continued)

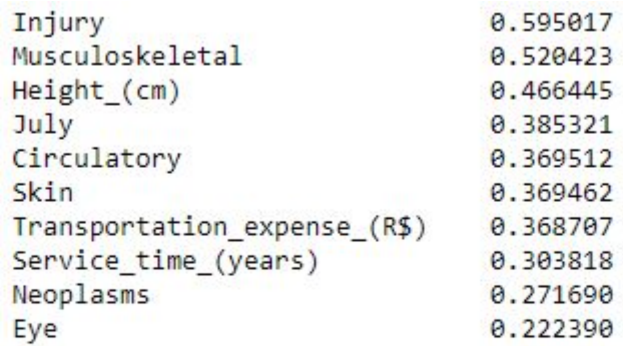


```
scaler = preprocessing.StandardScaler()

cls = linear_model.LogisticRegression()

cls.fit(X_train, y_train)
y_pred = cls.predict(X_test)

X_train = pd.DataFrame(scaler.fit_transform(X_train), columns=X_train.columns)
cls.fit(X_train, y_train)
coefs = pd.Series(cls.coef_[0], index=X_train.columns)
print (coefs.sort_values(ascending = False))
```



```
Injury                          0.595017
Musculoskeletal                 0.520423
Height_(cm)                     0.466445
July                            0.385321
Circulatory                     0.369512
Skin                            0.369462
Transportation_expense_(R$)     0.368707
Service_time_(years)            0.303818
Neoplasms                       0.271690
Eye                             0.222390
```

Workplace Absenteeism - Conclusions

Regression analyses showed that injury and musculoskeletal issues are the likeliest predictors of an employee missing 8 or more hours of work. Given the nature of courier work, one could expect courier employees to miss work due to these issues, though it is possible that chance plays a role in making these two issues the biggest.

Even if injuries and musculoskeletal issues being the biggest reasons for missing work is due to chance, steps can be taken to reduce the likelihood of employees getting injuries or musculoskeletal issues. Employers can issue guidelines to minimize the risk of injuries such as limits on what one person can carry at once and also minimize ergonomic hazards. Employees can exercise, stretch and make sure to get enough sleep to lower their chances of getting hurt. Public health officials can make the general public more aware of injuries on the job.

Workplace Absenteeism - Future Analyses

Courier is one type of industry, and this is one specific company, so data with the same factors from other courier companies and from other industries such as hospitality and tourism, agriculture, manufacturing, scientific research, and government employment would help to see if the industries have similar causes of workplace absenteeism, and also for public officials to more effectively target different industries in a large, diverse economy. Also, data from a longer time frame than just the 3 years in this data set could give more useful information.