Financial Statistics & Risk Management

Absenteeism at Work – Analysis & Prediction

Rutu Desai | Shubham Kokane

Summary / Abstract

Absenteeism is any failure to report for or remain at work as scheduled, regardless of the reason. This project aims to confirm if there is an absence problem among employees and to assess the impact it has on the organization. Absenteeism in the workplace is a disaster that stifles productivity and results.

We did data preprocessing, exploratory data analysis, and then feature engineering. Random Forest, Decision Tree, Logistic Regression, KNN, SVC classification models were used to make predictions.

Problem

Due to its unexpected nature, absenteeism is a severe workplace problem and an expensive occurrence for both employers and employees. You might think that the employee should be held accountable for taking a leave at all times, but this isn't always the case. There are many additional aspects to consider, such as the daily workload and commuting costs.

We have 19 explanatory variables that describe practically every aspect of an employee taking a day off. We can forecast employee absence hours by creating a relationship between the target variable and the explanatory variables. In total, the absenteeism at work dataset has 740 samples.

Data

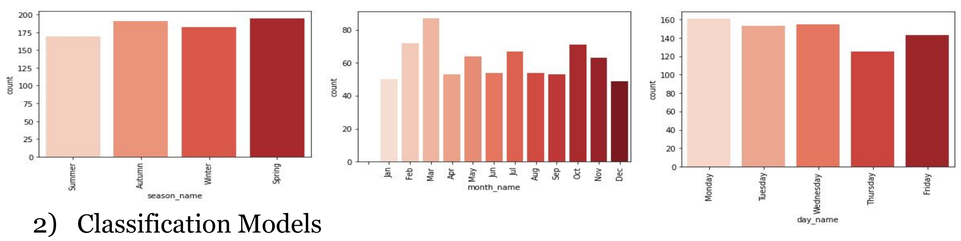
Every employee in this dataset has a unique id, and each leave they take is described by 19 predictors, including categorical variables such as reason for absence, education, season, smoker, etc. and discrete numerical predictors such as Transportation Expense, Workload, Body Mass Index, etc.

The target variable is Absenteeism time in hours, depicting the amount of time the employee takes a leave on a particular day, which we are trying to predict. The target variable had some null values that were handled during the preprocessing stage by calculating mean values of that column.

Methodology

1) Exploratory Data Analysis and Processing

1.1) Plotted the correlation heat map to see how different variables correlate with the target variable.
1.2) Found different categorical and numerical attributes, to further understand the trends in both the categories.
1.3) We were able to identify some issues in the dataset like there were more than 12 months, some employees were never absent in 3 years of time span.
1.4) Cleaned the dataset and made it consistent accordingly to make it useful for modeling.

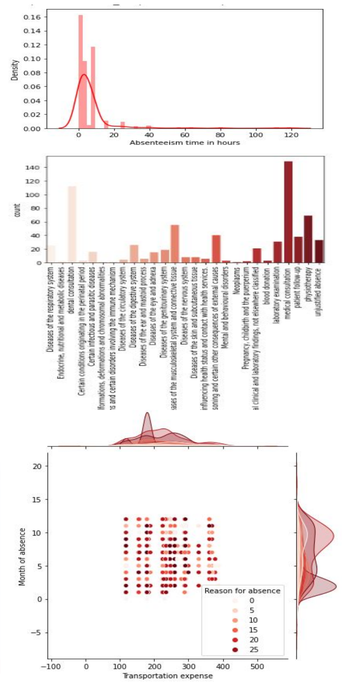


2) Classification Models

2.1) Random Forest 2.2) Decision Tree 2.3) Logistic Regression
2.4) K-Nearest Nearest Neighbors 2.5) Support Vector Machine (SVC)

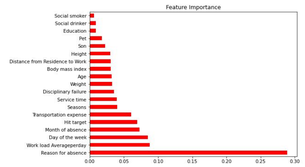
Evaluated model prediction accuracy using different scores like Precision, Recall and F1- Scores. Used Cross Validation for parameter tuning
Results Table:

Algorithm	Accuracy	Accuracy(Hyper parameter Tuning)		Precision, Recall, F1-Score			Precision, Recall, F1-Score(Hyper parameter Tuning)					
		Random Search CV	Grid Search CV				Random Search CV			Grid Search CV		
				Precision-Score	Recall Score	F1-Score	Precision-Score	Recall-Score	F1-Score	Precision-Score	Recall-Score	F1-Score
Random Forest(Without Age Range)	0.73			0.71	0.73	0.72						
Random Forest	0.74	0.73	0.74	0.72	0.74	0.73	0.72	0.74	0.73	0.72	0.74	0.73
Decision Tree	0.68	0.64	0.77	0.71	0.67	0.69	0.70	0.69	0.69	0.70	0.69	0.69
Logistics Regression	0.71	0.71	0.71	0.66	0.71	0.68	0.66	0.71	0.68	0.66	0.71	0.68
KNN Classification	0.69	0.66	0.67	0.69	0.69	0.67	0.69	0.69	0.67	0.69	0.69	0.67
SVM	0.63	0.56	0.56	0.62	0.63	0.53	0.62	0.63	0.54	0.62	0.63	0.54



Findings and Evaluation

1. 'Reason for absence,' 'Work load per day,' 'Month of absence,' and 'Week of absence' were shown to be the most important features/variables in defining employee absenteeism.
2. Most of the absences were taken during Spring Season and also on Mondays of the week.
3. To get better classification results, we divided the hours in 4 classes. (half day, full day, 2 days and 2 days or more)
4. The model with all features actually produces better classifying results than eliminating some features after using 'statsmodel.api', 'AIC' score and 'p-values(significant 0.05)' as threshold.
5. We tried Hyper parameter tuning with Randomized Search and Grid Search Cross Validation and the results were a bit better with specific parameters.
6. Then we tried different classification models with feature engineering using 'Forward Selection' and 'Backward Elimination' approaches.
7. With Forward Selection , we got better results when we used sequential feature selection method. It helped us reduce number of variables and showed better results.



Algorithm	Accuracy	Number of Features
Random Forest Classification	0.78	11
Decision Tree Classification	0.77	17
Logistic Regression	0.70	18
KNN Classification	0.78	5
SVC	0.69	5

Conclusions

According to our research, Random Forest Classification and KNN classification with 11 and 5 features respectively using Forward Selection gave us the best results. Also, because the target value of 'absenteeism in hours' is influenced by a variety of parameters, feature selection was a crucial step for us.

Future Scope of our project is to use more algorithms for regression and classifying the target variable using other feature selection methods to get better results and compare them with hyper parameter tuning. We can also look into how different approaches to classification of target variable can affect the accuracies.