OPIM 5604 – Predictive Modeling

Professor Eigo

Absenteeism at Work

Project #2

December 2, 2020

Team 3

Table of Contents

Executive Summary

In the courier industry, competition is fierce, and companies must rapidly innovate to keep up with their competitors and maintain customer satisfaction. The major factor in both is labor. With absenteeism comes a disruption in productivity and in turn disrupts profitability. The longer the absence of a worker, the larger the impact. This paper aims to help companies in the courier industry discover more insight into the factors that cause absenteeism by analyzing what factors influence the duration of a single absenteeism event. We used the SEMMA process to help us explore and analyze the dataset, as well as preprocess our data and build our predictive models.

First, we preprocessed our dataset by sampling, exploring, and modifying it. The major conclusions we discovered during this process were: ‘Reason for absence’ and ‘absenteeism time in hours’ had the strongest relationship. We needed to modify our target variable to fit our purposes. As the longer the absence the greater the disruption, we transformed our target variable into 0-8 and > 8 hours and classed it as a categorical variable. A validation column was also created to oversample > 8 hours period. Then we built our seven predictive models: Regression, Decision Tree, Neural Network, Discriminant Analysis, KNN, Naive Bayes, and Ensemble. Lastly, we assessed our predictive models on their misclassification rate and RMSE.

Based on each model’s misclassification rate, RASE, RSquare, and RMSE, we concluded that our Ensemble model predicted the duration of a single absenteeism event with the most accuracy. We recommend that courier companies do the following: (1) Establish attendance policies and attendance tracking system (Madlinger, Grace). (2) Evaluate and update health policies to include coverage that is relevant to employees (Madlinger, Grace). (3) Screen employees for variables related to absenteeism and continually monitor employee absenteeism. (4) Establish a system to reward employees who do not participate in absenteeism and enable other employees to improve. (5) Organize orientation programs highlighting the consequences of absenteeism. (6) Create flexible working options for the employees who work desk jobs and have a need to work from home.

Problem Statement

Every workplace deals with absenteeism. It is an unavoidable fact of life that planned and unplanned events will and can happen. When an employee is habitually absent from work it is costly. Not only does it affect the number of packages delivered and delivery times which in turn affects competition and customer satisfaction, but it can also increase the stress levels of other employees and affect their overall morale and their health. Thus, leading to a cycle of absenteeism and decreased productivity. We set out to find what are the factors and how the company might mitigate them.

Methodology

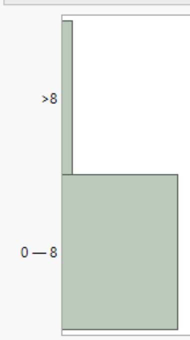
SAMPLE

The dataset used for our predictive modeling project is the Absenteeism at Work dataset from UCI Machine Learning Repository's archive of clean tabular datasets. The dataset was collected from a courier service in Brazil by PhD students at the Universidade Nove de Julho (UCI Machine Learning Repository: Absenteeism at Work Data Set). It contains 21 columns and 740 rows describing instances of absenteeism, general information about employees, and reasons for absenteeism recorded over the course of 3 years (UCI Machine Learning Repository: Absenteeism at Work Data Set).

EXPLORE

In this project, we are looking at 'Absenteeism in Hours' (target variable) and which predictive variables influence the duration of a single absentee event. Early on we found that rather than looking at the 38 employees (the dataset would be too small) we needed to look at each absentee event as unique (700 absentee events). When we looked at the distribution of 'Absenteeism in Hours' we found that it would be best to keep the data as it was. We determined that in our modification steps we needed to split it into two categories of less than 1 workday and more than 1 workday to predict the greatest impact. Based on the distribution we knew that more than 1 workday would be a rare event as the average was 3 hours.

Absenteeism time in hours Binned 2

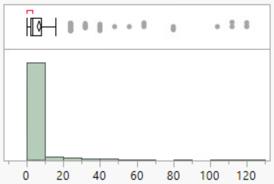


Frequencies

Level	Count	Prob
0 — 8	677	0.91486
>8	63	0.08514
Total	740	1.00000

Distributions

Absenteeism time in hours



Quantiles

100.0%	maximum	120
99.5%		114.36
97.5%		40
90.0%		8
75.0%	quartile	8
50.0%	median	3
25.0%	quartile	2
10.0%		1
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	6.9243243
Std Dev	13.330998
Std Err Mean	0.4900572
Upper 95% Mean	7.8863945
Lower 95% Mean	5.9622541
N	740

After performing ‘Multivariate Correlations’ on the continuous predictor variables we found that the most significant variables to ‘Absenteeism in Hours’ were: ‘Reason for Absence’ 17%, ‘Height’ 14%, ‘Day of the Week’ 12%, ‘Disciplinary Failure’ 12%, and ‘Son’ 11%.

Multivariate

Correlations

	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense
Reason for absence	1.0000	-0.0839	0.1163	-0.1179	-0.1194
Month of absence	-0.0839	1.0000	-0.0065	0.4078	0.1375
Day of the week	0.1163	-0.0065	1.0000	0.0465	0.0340
Seasons	-0.1179	0.4078	0.0465	1.0000	0.0370
Transportation expense	-0.1194	0.1375	0.0340	0.0370	1.0000
Distance from Residence to Work	0.1618	-0.0039	0.1180	-0.0631	0.2622
Service time	0.0484	-0.0629	0.0213	-0.0109	-0.3499
Age	-0.0786	-0.0015	0.0045	-0.0121	-0.2275
Work load Average/day	-0.1235	-0.1700	0.0156	0.1504	0.0054
Work load Average/day Binned	-0.1195	-0.1261	0.0056	0.1945	0.0332
Hit target	0.0889	-0.4605	0.0310	-0.0612	-0.0802
Disciplinary failure	-0.5451	0.1079	-0.0151	0.1518	0.1092
Education	-0.0474	-0.0661	0.0585	-0.0030	-0.0551
Son	-0.0554	0.0790	0.0981	0.0470	0.3830
Social drinker	0.0654	0.0562	0.0418	-0.0460	0.1451
Social smoker	-0.1157	-0.0386	0.0132	-0.0487	0.0444
Pet	-0.0559	0.0478	-0.0289	0.0124	0.4001
Weight	-0.0003	0.0233	-0.1290	-0.0263	-0.2074
Height	-0.0793	-0.0689	-0.0821	-0.0337	-0.1945
Absenteeism time in hours	-0.1731	0.0243	-0.1244	-0.0056	0.0276

	Distance from Residence to Work	Service time	Age	Work load Average/day	Work loday Binned	lay Binned	Hi t target
Reason for absence	0.1618	0.0484	-0.0786			-0.1195	0.0889
Month of absence	-0.0039	-0.0629	-0.0015	-0.1700		-0.1261	-0.4605
Day of the week	0.1180	0.0213	0.0045	0.0156		0.0056	0.0310
Seasons	-0.0631	-0.0109	-0.0121	0.1504		0.1945	-0.0612
Transportation expense	0.2622	-0.3499	-0.2275	0.0054		0.0332	-0.0802
Distance from Residence to Work	1.0000	0.1317	-0.1459	-0.0687		-0.0500	-0.0139
Service time	0.1317	1.0000	0.6710	-0.0007		0.0140	-0.0078
Age	-0.1459	0.6710	1.0000	-0.0394		-0.0315	-0.0392
Work load Average/day	-0.0687	-0.0007	-0.0394	1.0000		0.9599	-0.0894
Work load Average/day Binned	-0.0500	0.0140	-0.0315	0.9599		1.0000	-0.1282
Hit target	-0.0139	-0.0078	-0.0392	-0.0894		-0.1282	1.0000
Disciplinary failure	-0.0565	-0.0002	0.1043	0.0290		0.0438	-0.1480
Education	-0.2596	-0.2130	-0.2219	-0.0750		-0.1013	0.1011
Son	0.0542	-0.0471	0.0570	0.0278		0.0347	-0.0141
Social drinker	0.4522	0.3531	0.2132	-0.0337		-0.0221	-0.1025
Social smoker	-0.0754	0.0724	0.1217	0.0310		0.0178	0.0513
Pet	0.2059	-0.4403	-0.2312	0.0071		0.0148	0.0072
Weight	-0.0479	0.4560	0.4187	-0.0385		-0.0393	-0.0449
Height	-0.3534	-0.0531	-0.0630	0.1033		0.0718	0.0933
Absenteeism time in hours	-0.0884	0.0190	0.0658	0.0247		0.0197	0.0267

Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Absenteeism time in hours
-0.5451	-0.0474	-0.0554	0.0654	-0.1157	-0.0559	-0.0003	-0.0793	-0.1731
0.1079	-0.0661	0.0790	0.0562	-0.0386	0.0478	0.0233	-0.0689	0.0243
-0.0151	0.0585	0.0981	0.0418	0.0132	-0.0289	-0.1290	-0.0821	-0.1244
0.1518	-0.0030	0.0470	-0.0460	-0.0487	0.0124	-0.0263	-0.0337	-0.0056
0.1092	-0.0551	0.3830	0.1451	0.0444	0.4001	-0.2074	-0.1945	0.0276
-0.0565	-0.2596	0.0542	0.4522	-0.0754	0.2059	-0.0479	-0.3534	-0.0884
-0.0002	-0.2130	-0.0471	0.3531	0.0724	-0.4403	0.4560	-0.0531	0.0190
0.1043	-0.2219	0.0570	0.2132	0.1217	-0.2312	0.4187	-0.0630	0.0658
0.0290	-0.0750	0.0278	-0.0337	0.0310	0.0071	-0.0385	0.1033	0.0247
0.0438	-0.1013	0.0347	-0.0221	0.0178	0.0148	-0.0393	0.0718	0.0197
-0.1480	0.1011	-0.0141	-0.1025	0.0513	0.0072	-0.0449	0.0933	0.0267
1.0000	-0.0593	0.0721	0.0518	0.1167	0.0189	0.0722	-0.0105	-0.1242
-0.0593	1.0000	-0.1886	-0.4200	0.0327	-0.0536	-0.3006	0.1010	-0.0462
0.0721	-0.1886	1.0000	0.2064	0.1561	0.1089	-0.1396	-0.0142	0.1138
0.0518	-0.4200	0.2064	1.0000	-0.1117	-0.1228	0.3787	0.1700	0.0651
0.1167	0.0327	0.1561	-0.1117	1.0000	0.1054	-0.1985	0.0033	-0.0089
0.0189	-0.0536	0.1089	-0.1228	0.1054	1.0000	-0.1038	-0.1031	-0.0283
0.0722	-0.3006	-0.1396	0.3787	-0.1985	-0.1038	1.0000	0.3068	0.0158
-0.0105	0.1010	-0.0142	0.1700	0.0033	-0.1031	0.3068	1.0000	0.1444
-0.1242	-0.0462	0.1138	0.0651	-0.0089	-0.0283	0.0158	0.1444	1.0000

MODIFY

We performed the following treatments to our dataset: (1) We binned our target variable (Absenteeism Time in Hours) into 2 categories - 0 to 8 hours (1 workday or less) and > 8 hours. (more than 1 workday). We also reclassified it as a nominal variable as it would aid us in building models that predicted events of > 8 hours, to better identify events that would cause the greatest impact. (2) We partitioned our data into 60% Training and 40% Validation in accordance with the size of the dataset. Since > 8 hours occurs less often than 0 to 8 hours, we also created an extra validation column that oversampled this event. (4) There were no missing values in this data, however, there were outliers in Age (no treatment - relevant to data and small in scale in terms of variance) and Height (transformed to normal quantile). (5) We decided to exclude BMI as it would be redundant to Weight and Height. (6) We binned Workload Average/Day into 40,000 range bins. The data was transformed from 38 unique values to 5 unique ranges.

Please see Appendix B: JMP Screenshots - Modify

MODEL

Regression Models

For our Regression models, we built both Logistic and Linear regression models with all our predictor variables. We determined that the best model of the two was the Logistic Regression model and decided to use this model for our final model comparison. The most influential variable in our logistic regression model was the Reason for Absence, as each category was both substantially influential on the prediction and differentiated from other categories. The only other truly noteworthy differentiation could be found in Education. Having secondary education makes having a long absentee event much less likely.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Regression Models

Decision Tree Model

For our best Decision Tree model, we used all the predictor variables and the binned target variable as the inputs. The resultant model had just 1 split at "reason for absence". Additionally, we used the validation column which oversamples our dataset to build our decision tree models. On further pruning and testing with additional splits, we were able to build our optimum tree model with 4 splits at ‘Reason for Absence’, ‘Social Drinker’, ‘Day of the Week’, and ‘Distance from Residence to Work’.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Decision Tree Model

Discriminant Analysis Model

The best Discriminant Analysis model we built used all predictive variables. With this model, we adjusted the prior probabilities to match our population and the cutoff to register calculated probabilities of taking the additional sick time of 0.3 or greater as “>8” events. However, we did not retain the adjusted cutoff for our ensemble model since that would disrupt the voting algorithm. The discriminant analysis model type was not a good fit for our data, since few of our predictors were normal, there were many outliers within the critical Reason for Absenteeism variable, and not all predictors were equally correlated with the target variable.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Discriminant Analysis Model

Neural Network Model

Our best Neural Network model used the following variables for the input layer: ‘Reason for Absence’, ‘Day of the Week’, ‘Distance from Residence’, ‘Social Drinker’, ‘Pet’, and ‘Height.’ We also used 1 hidden layer with 3 nodes: 1 Tanh, 1 Linear, and 1 Gaussian. The model is combining the input information and captures complex relationships between the outcome and predictors. There are multiple iterations done to find weights that give the best results.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Neural Network Model

KNN Model

The KNN model can be used for both categorical and continuous variables, so we build 2 models with different outcomes. For continuous target variables, we first use all the predictive variables to build the model. To reduce dimensionality, we use PCs to build the model again. For categorical target variables, we still use all the predictive variables to build the model first and then use PCs to build the model again. We determined that the best model for our model comparison was the model built with categorical variables.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - KNN Model

Naive Bayes

The Naive Bayes model looks at the classification of existing records to classify a new record. The best model that we built performed fairly well using all the categorical predictive variables. We changed the variables ‘Pet’ and ‘Son’ to nominal and used these columns, along with the rest of the categorical variables in the data set, as the predictors for this model to predict the binned ‘Absenteeism time in hours.’ Since there were no continuous variables that were significant predictors for our model, it wasn’t necessary to add any to this one.

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Naive Bayes Model

Ensemble

Our Ensemble model includes the results from Logistic, Neural Network, Naïve Bayes, KNN, and Decision Tree models. We used the ‘Model Averaging’ tool to create our ensemble model. Below are the formulas for our 0 to 8 and > 8 hours predicted probabilities, generated by using ‘Model Averaging’ within JMP, which simply averages the fits for the various models.

$$\text{If}\begin{cases}\text{Format}(\textit{Validation}) == \text{"Validation"} \Rightarrow \dfrac{\left(\textit{Prob[0 — 8]} + \textit{Probability(Absenteeism time in hours Binned 2=0 — 8)} + \textit{Prob(Absenteeism time in hours Binned 2==0 — 8)}\right)}{3} \\ \text{Format}(\textit{Validation}) == \text{"Training"} \Rightarrow \dfrac{\left(\textit{Prob[0 — 8]} + \textit{Probability(Absenteeism time in hours Binned 2=0 — 8)} + \textit{Prob(Absenteeism time in hours Binned 2==0 — 8)}\right)}{3} \\ \text{else} \Rightarrow .\end{cases}$$

$$\text{If}\begin{cases}\text{Format}(\textit{Validation}) == \text{"Validation"} \Rightarrow \dfrac{\left(\textit{Prob[>8]} + \textit{Probability(Absenteeism time in hours Binned 2=>8)} + \textit{Prob(Absenteeism time in hours Binned 2==>8)}\right)}{3} \\ \text{Format}(\textit{Validation}) == \text{"Training"} \Rightarrow \dfrac{\left(\textit{Prob[>8]} + \textit{Probability(Absenteeism time in hours Binned 2=>8)} + \textit{Prob(Absenteeism time in hours Binned 2==>8)}\right)}{3} \\ \text{else} \Rightarrow .\end{cases}$$

Additionally, we used the voting method to formulate our ensemble model classifications, classifying records with the category that was predicted by at least three of our five major models.

	nteeism time in hours	Absenteeism time in hours ...	Validation	Validation Stratified by Absenteeism time in ...	Logistic Regression Classification	Neural Model Classification	Naive Bayes Predictions	Decision Tree Classification - ...	KNN Classification	Absenteeism Binned 0 — 8 Avg Predictor By Validation	Absenteeism Binned 2 >8 Avg Predictor By Validation	Ensemble Model
377	3	0 — 8	Validation	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9724200595	0.0275799405	0 — 8
378	8	0 — 8	Training	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.8722341293	0.1277658707	0 — 8
379	8	0 — 8	Validation	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9417591041	0.0582408959	0 — 8
380	3	0 — 8	Training	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9908629273	0.0091370727	0 — 8
381	8	0 — 8	Training	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9936611265	0.0063388735	0 — 8
382	3	0 — 8	Validation	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9506423219	0.0493576781	0 — 8
383	2	0 — 8	Training	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9822329332	0.0177670668	0 — 8
384	2	0 — 8	Training	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9816103155	0.0183896845	0 — 8
385	16	>8	Training	Training	>8	0 — 8	0 — 8	>8	0 — 8	0.3618098543	0.6381901457	0 — 8
386	3	0 — 8	Validation	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9722729047	0.0277270953	0 — 8
387	3	0 — 8	Training	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9908271403	0.0091728597	0 — 8
388	24	>8	Training	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.8398777428	0.1601222572	0 — 8
389	3	0 — 8	Validation	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9722009435	0.0277990565	0 — 8
390	3	0 — 8	Training	Validation	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9908002826	0.0091997174	0 — 8
391	8	0 — 8	Validation	Training	0 — 8	0 — 8	0 — 8	0 — 8	0 — 8	0.9910405342	0.0089594658	0 — 8
392	16	>8	Training	Training	>8	0 — 8	>8	>8	0 — 8	0.2857006645	0.7142993355	>8

Please see Appendix A: Predictive Variables and Appendix B: JMP Screenshots - Model - Ensemble Model

ASSESS

After building and determining the best model in each model type we needed to determine the best model overall. We decided on measuring each model’s performance based on their misclassification rate, RSquare, RASE, and RMSE in the model comparison application in JMP.

Results

We determined based on the performance of the ‘Model Comparison’ application that the best model overall was the Ensemble Model. We based our decision on the Validation misclassification rate and RMSE. The ensemble model has a misclassification rate of 0.0473, RSquare of 0.4010, an RMSE of 0.2089 and an overall accuracy of 95.27%. The ensemble model has a relatively higher overall accuracy, and it classifies 3.4% of the validation records as “>8” hours of absenteeism and has an accuracy of 70% for “>8” predictions and 95% for “0 - 8” predictions. The logistic regression model on the other hand classifies 6.42% of the validation records as “>8” but has a relatively lower accuracy of 57.9% for “>8” predictions. The accuracy of “0 - 8” predictions is 93.5%, which is promising. The ensemble model captures sufficient variance and inherent patterns from the training partition while remaining relatively less complex and performing well on the validation partition. The Naïve Bayes model classifies 8.45% of the validation partition as “>8” but has a relatively lower accuracy of 52% for “>8” predictions. The Ensemble model has a better rate of classification and a higher accuracy of predictions for both (0 - 8, >8) classes of the categorical target variable.

Absenteeism_WithEnsembleModel - 3 - Model Comparison 6 - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for Absenteeism time in hours Binned 2

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		-0.105	-0.147	0.346	0.2546	0.0975	0.0766	444
Neural		0.2059	0.2599	0.2486	0.2632	0.1237	0.0901	444
Naive Bayes		.	.	.	.	.	0.0878	444
Partition		0.3851	0.4605	0.1925	0.2347	0.1108	0.0676	444
K Nearest Neighbors		.	.	.	.	.	0.0946	444
Model Averaged		0.3936	0.4694	0.1898	0.2355	0.1107	0.0743	444

Model Comparison Validation=Validation

Predictors

Measures of Fit for Absenteeism time in hours Binned 2

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		-0.821	-1.305	0.4664	0.2254	0.0775	0.0608	296
Neural		0.1943	0.2363	0.2063	0.2324	0.1099	0.0709	296
Naive Bayes		.	.	.	.	.	0.0676	296
Partition		0.2399	0.2884	0.1947	0.2313	0.1054	0.0676	296
K Nearest Neighbors		.	.	.	.	.	0.0709	296
Model Averaged		0.3421	0.4010	0.1685	0.2089	0.0976	0.0473	296

Conclusions and Recommendations

Based on our findings from our preprocessing steps and our final model, we have drawn several conclusions from the Absenteeism at Work dataset. ‘Reason for Absence’, ‘Day of the Week’, ‘Disciplinary Failure’ and ‘Son’ had the most significant relationships with ‘Absenteeism in Hours.’ With this information, we found out the following: (1) The most common ‘Reasons for Absence’ are (23) Blood Donation 20%, (28) Dental Consultation 15%, and (27) Physiotherapy 9%. (2) Employees are more likely to be absent on (2) Monday 21% than on (5) Friday 16%. (3) Those who never received disciplinary failure are more likely to be absent (95%) than those who haven't (5%). Disciplinary action seems to be an effective tool to motivate employees’ punctuality and presence in the workplace. (4) On average employees have at least 1 son. (5) Additionally, it turns out that employees with a higher level of education are more likely to be present to work and on time. (6) Further, it seems that employees who are between the age of 30 to 45 years are late to work more often. This could be because they have kids and are required to attend school events or take them to doctor’s appointments etc. They are probably tardy more owing to a more active family life. (7) Employees with a higher percentage of ‘hit target’ are usually late to work compared to employees who have not been able to meet a high percentage of their target.

Please see Appendix B: JMP Screenshots - Model Insights

Based on our findings and conclusions in this project, we made the following recommendations: (1) Establish attendance policies and attendance tracking systems - addressing childcare and appropriate absences (Madlinger, Grace). (2) Evaluate and update health policies to include coverage that is relevant to employees, such as physiotherapy and dentistry (Madlinger, Grace). (3) Screen employees for variables related to absenteeism and continually monitoring employee absenteeism. (4) Establish a system to reward employees who do not participate in absenteeism and enable other employees to improve. (5) Organize orientation programs highlighting the consequences of absenteeism. (6) Create flexible working options for the employees who work desk jobs and have a need to work from home, especially those with children. (7) The company should think about engaging high performing employees who have greater ‘Hit Rates’ in challenging activities and give them significant tasks to work on and opportunities to master their skills otherwise tardiness can potentially creep in.

The success of a courier company is based on the rate of innovation compared to competition and customer satisfaction. Both hinder the productivity of the employees. A disruption in labor affects production, profit, and morale. These recommendations ensure employee loyalty as it recognizes employees' needs while limiting future production disruptions. Consistent and loyal labor allows for consistent productivity and profitability.

References

Data Mining for Business Analytics. Wiley, 2016.

Madlinger, Grace. “The 6-Step Process for Dealing with Employee Absenteeism.” When I Work, 20 Mar. 2018, wheniwork.com/blog/how-to-deal-with-employee-absenteeism.

UCI Machine Learning Repository: Absenteeism at Work Data Set. archive.ics.uci.edu/ml/datasets/Absenteeism+at+work.

Wakabayashi, Daisuke & Sheera Frenkel. “Parents Got More Time Off. Then the Backlash Started.” New York Times. September 5, 2020. https://www.nytimes.com/2020/09/05/technology/parents-time-off-backlash.html

Appendix

Appendix A: Predictive Variables

Serial No#	Attribute	Possible values	Data Type	Attribute Information
1	ID		Nominal	ID corresponds to the Identification Codes for individual employees at the Courier Company
2	Reason For Absence	28 Unique Values	Nominal	This attribute represents the medical reasons given by employees while taking leaves. The concerned courier company recognizes 28 reasons for absence, out of which 21 are attested by the International Code of Diseases (ICD) and 7 are categories without ICD attestation.
3	Month Of Absence	1 To 12	Categorical/ Nominal	Corresponds to the months of a year, with 1 representing January; 2 representing February and so on. Each row relative to this variable represents the month when an employee was absent from work.

4	Day Of The Week	Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)	Categorical/ Nominal	This column describes days when individual employees were absent, with each instance (corresponding row) representing a distinct day.
5	Seasons	Summer (1), Autumn (2), Winter (3), Spring (4)	Categorical/ Nominal	This column represents seasons.
6	Transportation Expense	Range: 118 - 388, Mean: 221.33, Median: 225	Continuous	Continuous variable describing the transportation expenses.
7	Distance From Residence To Work		Continuous	Continuous variable describing the distance from residence to the workplace of individual employees in Kilometers.
8	Service Time		Continuous	The documented duration of service in the week when the concerned employee took leave from work.
9	Age		Continuous	Age of the employee.
10	Workload Average/Day [Binned @ 40,000]		Categorical/ Nominal	The average workload per day.
11	Hit Target		Continuous	% of employee's achievement of periodic goals
12	Disciplinary Failure	Yes=1 No=0	Categorical/Nominal	Whether or not the employee faced disciplinary action.
13	Education	High School (1), Graduate (2), Postgraduate (3), Master and Doctor (4)	Ordinal	The level of education the employee has.
14	Son		Continuous	# of children the employee has.
15	Social Drinker	Yes=1 No=0	Categorical/Nominal	Whether or not the employee is a social drinker.
16	Social Smoker	Yes=1 No=0	Categorical/Nominal	Whether or not the employee is a social smoker.

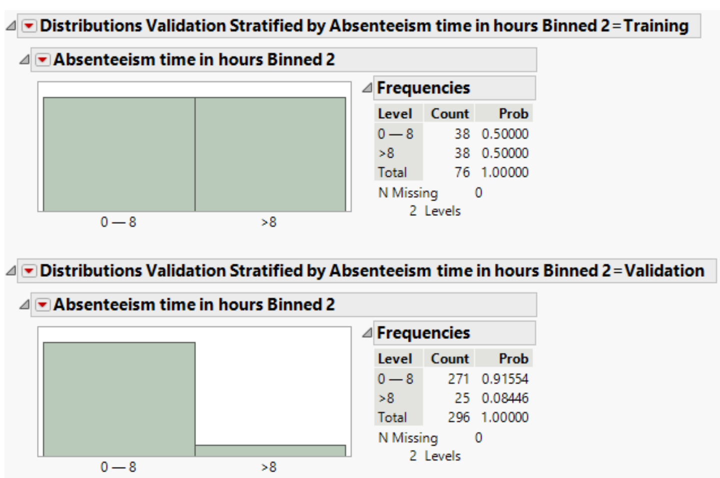
17	Pet		Continuous	# of Pets
18	Weight		Continuous	Weight of Employee in kg
19	Height [Normal Quantile]		Continuous	Height of Employee in cm
20	Body Mass Index [Excluded]		Continuous	BMI of Employee
21	**Absenteeism Time in Hours (Target Variable) [Binned 0-8 and >8]**		Continuous	# of Absenteeism Hours

Appendix B: JMP Screenshots

Sample

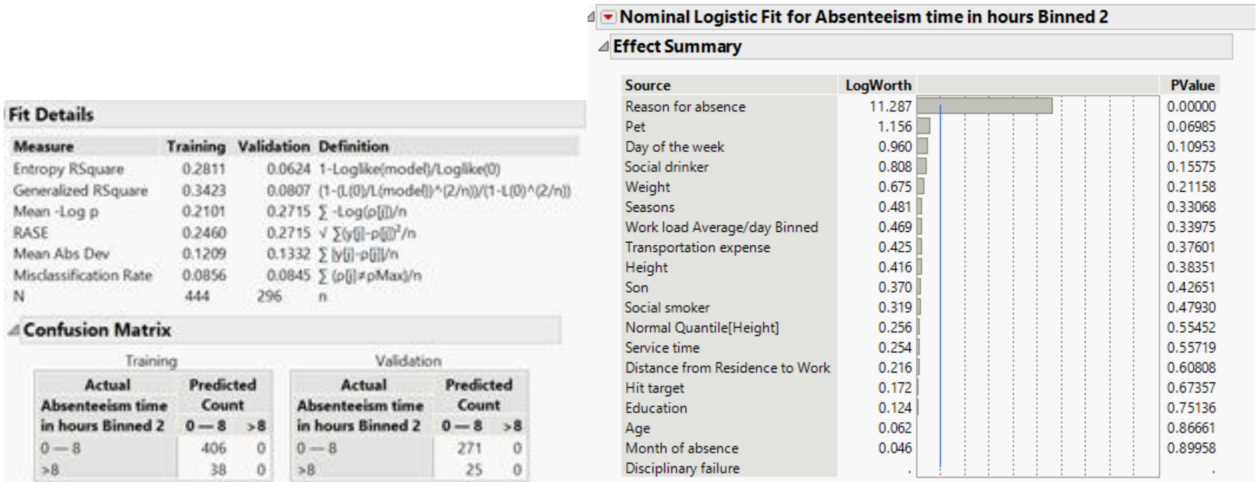
Explore

Modify



Model

Regression Models



Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.2811	0.0624	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3423	0.0807	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2101	0.2715	$\sum$ -Log(ρ[j])/n
RASE	0.2460	0.2715	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1209	0.1332	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0856	0.0845	$\sum$ (ρ[j]≠ρMax)/n
N	444	296	n

Confusion Matrix

Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	406	0
>8	38	0

Validation

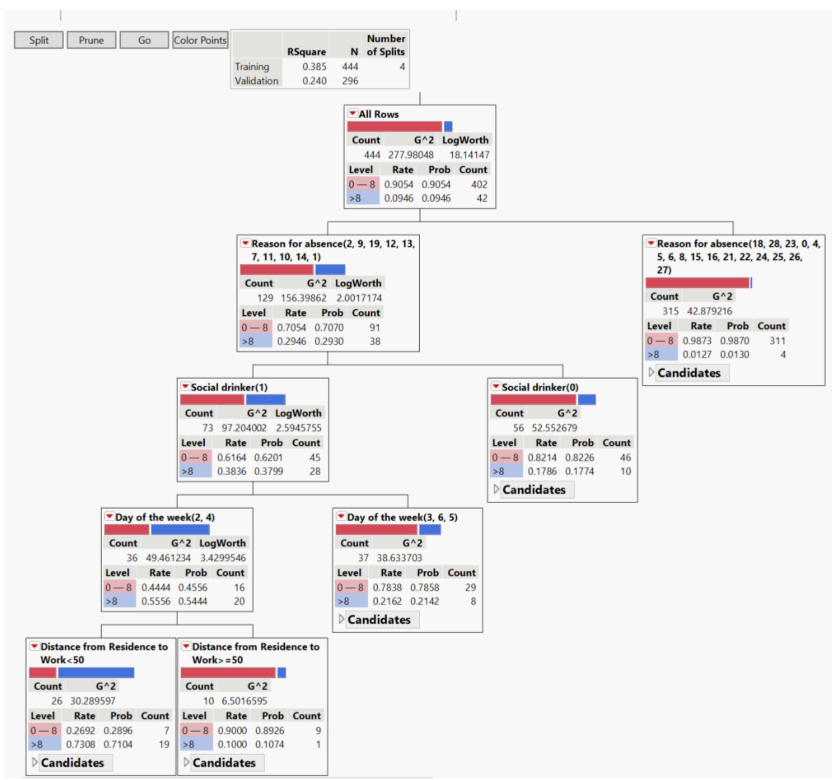
Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	271	0
>8	25	0

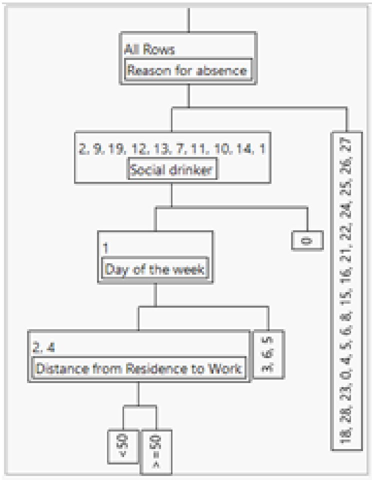
Nominal Logistic Fit for Absenteeism time in hours Binned 2

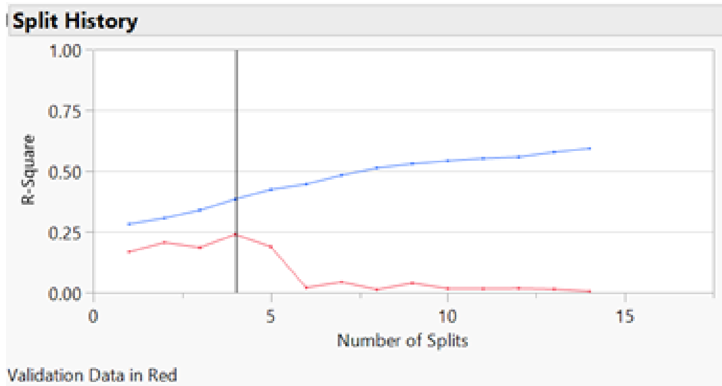
Effect Summary

Source	LogWorth	PValue
Reason for absence	11.287	0.00000
Pet	1.156	0.06985
Day of the week	0.960	0.10953
Social drinker	0.808	0.15575
Weight	0.675	0.21158
Seasons	0.481	0.33068
Work load Average/day Binned	0.469	0.33975
Transportation expense	0.425	0.37601
Height	0.416	0.38351
Son	0.370	0.42651
Social smoker	0.319	0.47930
Normal Quantile[Height]	0.256	0.55452
Service time	0.254	0.55719
Distance from Residence to Work	0.216	0.60808
Hit target	0.172	0.67357
Education	0.124	0.75136
Age	0.062	0.86661
Month of absence	0.046	0.89958
Disciplinary failure	.	.

Decision Tree Model







Leaf Report

Response Prob

Leaf Label	0 — 8	>8
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(2, 4)&Distance from Residence to Work<50	0.2896	0.7104
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(2, 4)&Distance from Residence to Work>=50	0.8926	0.1074
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(3, 6, 5)	0.7858	0.2142
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(0)	0.8226	0.1774
Reason for absence(18, 28, 23, 0, 4, 5, 6, 8, 15, 16, 21, 22, 24, 25, 26, 27)	0.9870	0.0130

Response Counts

Leaf Label	0 — 8	>8
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(2, 4)&Distance from Residence to Work<50	7	19
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(2, 4)&Distance from Residence to Work>=50	9	1
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(1)&Day of the week(3, 6, 5)	29	8
Reason for absence(2, 9, 19, 12, 13, 7, 11, 10, 14, 1)&Social drinker(0)	46	10
Reason for absence(18, 28, 23, 0, 4, 5, 6, 8, 15, 16, 21, 22, 24, 25, 26, 27)	311	4

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3851	0.2399	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4605	0.2884	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1925	0.1947	∑ -Log(ρ[j])/n
RASE	0.2347	0.2313	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1108	0.1054	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0676	0.0676	∑ (ρ[j]≠ρMax)/n
N	444	296	n

Confusion Matrix

Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	395	7
>8	23	19

Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	268	7
>8	13	8

Boosted Tree Model

Boosted Tree for Absenteeism time in hours Binned 2

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.3894	0.1182	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4598	0.1506	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1784	0.2553	$\sum$ -Log(ρ[j])/n
RASE	0.2280	0.2653	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1144	0.1276	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0811	0.0845	$\sum$ (ρ[j]≠ρMax)/n
N	444	296	n

Confusion Matrix

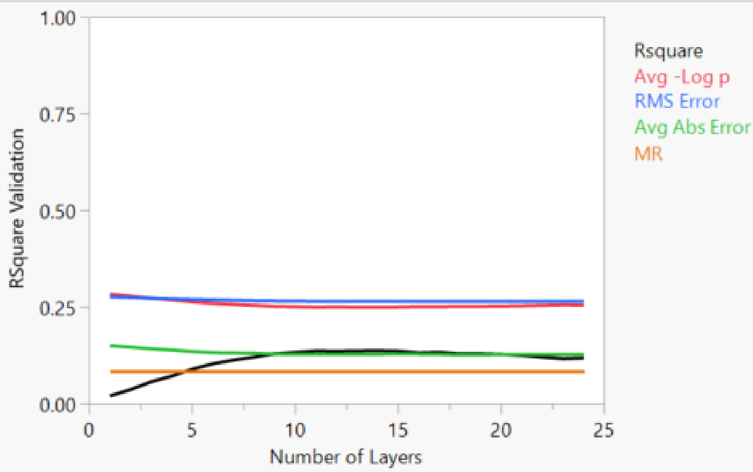
Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	Predicted Count >8
0 — 8	406	0
>8	36	2

Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	Predicted Count >8
0 — 8	271	0
>8	25	0

Cumulative Validation



Bootstrap Forest Model

Bootstrap Forest for Absenteeism time in hours Binned 2

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.3942	0.2096	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4700	0.2540	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1896	0.2024	$\sum$ -Log(ρ[j])/n
RASE	0.2368	0.2352	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1315	0.1326	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0946	0.0709	$\sum$ (ρ[j]≠ρMax)/n
N	444	296	n

Confusion Matrix

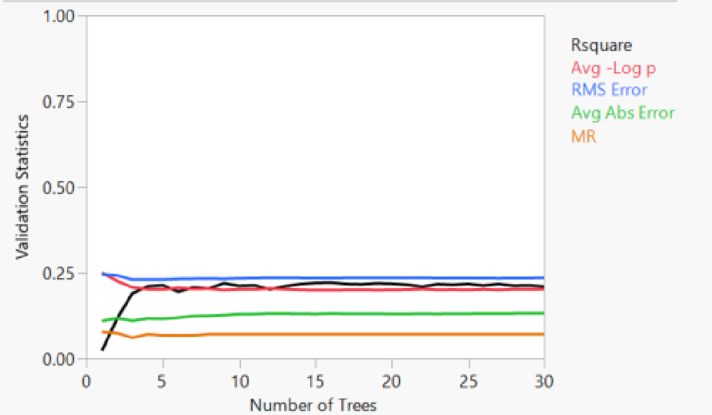
Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	402	0
>8	42	0

Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	275	0
>8	21	0

Cumulative Validation



Neural Network Model

Confusion Matrix

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	401	5
>8	30	8

Confusion Rates

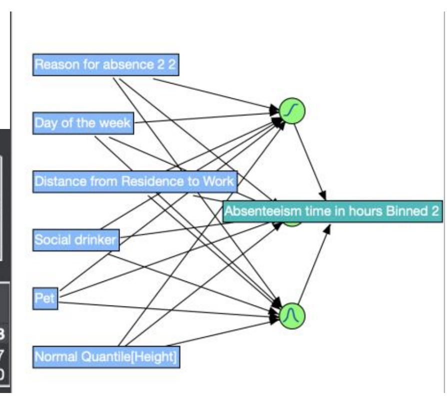
Actual Absenteeism time in hours Binned 2	Predicted Rate 0 — 8	>8
0 — 8	0.988	0.012
>8	0.789	0.211

Confusion Matrix

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	269	2
>8	22	3

Confusion Rates

Actual Absenteeism time in hours Binned 2	Predicted Rate 0 — 8	>8
0 — 8	0.993	0.007
>8	0.880	0.120



Estimates

Parameter	Estimate
H1_1:Reason for absence 2 2:1	-14.2851
H1_1:Reason for absence 2 2:2	-3.30474
H1_1:Reason for absence 2 2:3	6.460725
H1_1:Reason for absence 2 2:4	-1.51236
H1_1:Reason for absence 2 2:5	0.682486
H1_1:Reason for absence 2 2:6	2.209685
H1_1:Reason for absence 2 2:7	1.34471
H1_1:Reason for absence 2 2:8	-1.87216
H1_1:Reason for absence 2 2:9	-3.26155
H1_1:Reason for absence 2 2:10	-4.53345
H1_1:Reason for absence 2 2:11	-0.62883
H1_1:Reason for absence 2 2:12	-3.9237
H1_1:Reason for absence 2 2:13	-1.97881
H1_1:Reason for absence 2 2:14	-1.28169
H1_1:Reason for absence 2 2:15	-1.75089
H1_1:Reason for absence 2 2:16	3.2808
H1_1:Reason for absence 2 2:17	1.863819
H1_1:Reason for absence 2 2:18	1.972338
H1_1:Reason for absence 2 2:19	1.21902
H1_1:Reason for absence 2 2:20	3.59837
H1_1:Reason for absence 2 2:21	1.871631
H1_1:Reason for absence 2 2:22	2.607499
H1_1:Reason for absence 2 2:23	-0.61079
H1_1:Reason for absence 2 2:24	2.565152
H1_1:Reason for absence 2 2:25	-6.41511
H1_1:Reason for absence 2 2:26	-0.77668
H1_1:Reason for absence 2 2:27	-1.52495
H1_1:Day of the week:2	1.318696
H1_1:Day of the week:3	-0.52639
H1_1:Day of the week:4	1.090348
H1_1:Day of the week:5	-1.09912
H1_1:Distance from Residence to Work	-0.04639
H1_1:Social drinker:0	1.413925
H1_1:Pet	-1.03053
H1_1:Normal Quantile[Height]	0.210375
H1_1:Intercept	-35.7125
H1_2:Reason for absence 2 2:1	0.015831
H1_2:Reason for absence 2 2:2	3.564753
H1_2:Reason for absence 2 2:3	-0.96986

Parameter	Estimate
H1_2:Reason for absence 2 2:4	-2.17977
H1_2:Reason for absence 2 2:5	-4.98483
H1_2:Reason for absence 2 2:6	2.756113
H1_2:Reason for absence 2 2:7	2.016124
H1_2:Reason for absence 2 2:8	-0.01759
H1_2:Reason for absence 2 2:9	-3.7176
H1_2:Reason for absence 2 2:10	1.487438
H1_2:Reason for absence 2 2:11	1.934893
H1_2:Reason for absence 2 2:12	4.382212
H1_2:Reason for absence 2 2:13	4.18997
H1_2:Reason for absence 2 2:14	3.8986
H1_2:Reason for absence 2 2:15	-0.62342
H1_2:Reason for absence 2 2:16	-0.07642
H1_2:Reason for absence 2 2:17	0.363922
H1_2:Reason for absence 2 2:18	-3.05731
H1_2:Reason for absence 2 2:19	6.327649
H1_2:Reason for absence 2 2:20	-1.46667
H1_2:Reason for absence 2 2:21	0.543256
H1_2:Reason for absence 2 2:22	-1.0357
H1_2:Reason for absence 2 2:23	-2.16601
H1_2:Reason for absence 2 2:24	1.905656
H1_2:Reason for absence 2 2:25	-3.05335
H1_2:Reason for absence 2 2:26	-1.19391
H1_2:Reason for absence 2 2:27	-2.36777
H1_2:Day of the week:2	1.670182
H1_2:Day of the week:3	0.459382
H1_2:Day of the week:4	1.036869
H1_2:Day of the week:5	0.563631
H1_2:Distance from Residence to Work	0.011438
H1_2:Social drinker:0	-1.35921
H1_2:Pet	-1.13504
H1_2:Normal Quantile[Height]	0.274337
H1_2:Intercept	-44.4865
H1_3:Reason for absence 2 2:1	-6.32085
H1_3:Reason for absence 2 2:2	2.200769
H1_3:Reason for absence 2 2:3	-1.31403
H1_3:Reason for absence 2 2:4	4.272612
H1_3:Reason for absence 2 2:5	1.646305
H1_3:Reason for absence 2 2:6	-0.19575
H1_3:Reason for absence 2 2:7	1.702996
H1_3:Reason for absence 2 2:8	-2.96869
H1_3:Reason for absence 2 2:9	1.140195
H1_3:Reason for absence 2 2:10	2.675935
H1_3:Reason for absence 2 2:11	2.33346

Estimates

Parameter	Estimate
H1_3:Reason for absence 2 2:11	2.33346
H1_3:Reason for absence 2 2:12	-3.34867
H1_3:Reason for absence 2 2:13	-2.06448
H1_3:Reason for absence 2 2:14	-3.41744
H1_3:Reason for absence 2 2:15	0.715013
H1_3:Reason for absence 2 2:16	1.766398
H1_3:Reason for absence 2 2:17	1.754686
H1_3:Reason for absence 2 2:18	-2.34973
H1_3:Reason for absence 2 2:19	0.808534
H1_3:Reason for absence 2 2:20	1.478719
H1_3:Reason for absence 2 2:21	0.751471
H1_3:Reason for absence 2 2:22	2.324858
H1_3:Reason for absence 2 2:23	-3.71378
H1_3:Reason for absence 2 2:24	-0.35483
H1_3:Reason for absence 2 2:25	-0.39633
H1_3:Reason for absence 2 2:26	2.054207
H1_3:Reason for absence 2 2:27	1.346927
H1_3:Day of the week:2	-1.27412
H1_3:Day of the week:3	0.821133
H1_3:Day of the week:4	-1.8393
H1_3:Day of the week:5	-0.06275
H1_3:Distance from Residence to Work	-0.07143
H1_3:Social drinker:0	0.53498
H1_3:Pet	1.09672
H1_3:Normal Quantile[Height]	-0.03726
H1_3:Intercept	10.33484
Absenteeism time in hours Binned 2(0 — 8):H1_1	0.059161
Absenteeism time in hours Binned 2(0 — 8):H1_2	-0.34793
Absenteeism time in hours Binned 2(0 — 8):H1_3	-0.08071
Absenteeism time in hours Binned 2(0 — 8):Intercept	3.970315

Discriminant Analysis Model

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	444	112	25.2252	-0.8988	492.697
Validation	296	61	20.6081	-0.8394	

Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	321	85
>8	27	11

Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	224	47
>8	14	11

	DA .3 Cut-off	
Absenteeism time in hours Binned 2	>8	0-8
0 — 8	495	182
>8	59	4

Adjusted Cutoff

KNN Model

Training

K	Count	Misclassification Rate	Misclassifications
8	444	0.09009	40
9	444	0.09009	40
10	444	0.09009	40
11	444	0.08559	38 *
12	444	0.08559	38
13	444	0.08559	38
14	444	0.08559	38
15	444	0.08559	38
16	444	0.08559	38
17	444	0.08559	38

Validation

K	Count	Misclassification Rate	Misclassifications
6	293	0.09898	29
7	293	0.09898	29
8	293	0.09898	29
9	293	0.09215	27
10	293	0.08874	26
11	293	0.08532	25 *
12	293	0.08532	25
13	293	0.08532	25
14	293	0.08532	25
15	293	0.08532	25

Confusion Matrix for Best K=11

Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	406	0
>8	38	0

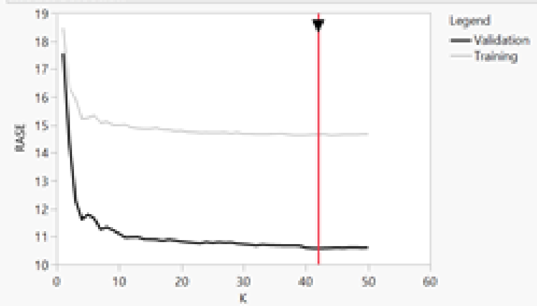
Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	268	0
>8	25	0

K Nearest Neighbors

Absenteeism time in hours

Model Selection



Training

K	Count	RSquare	RASE	SSE
41	444	0.01637	14.653	95332.1
42	444	0.01544	14.660	95422.9
43	444	0.01433	14.668	95529.7
44	444	0.01931	14.631	95047.3 *
45	444	0.01817	14.640	95157.8
46	444	0.01655	14.652	95314.4
47	444	0.01661	14.651	95309
48	444	0.01664	14.651	95305.7
49	444	0.01463	14.666	95501
50	444	0.01427	14.669	95535.6

Validation

K	Count	RSquare	RASE	SSE
41	293	0.04835	10.582	32808.3
42	293	0.04974	10.574	32760.2 *
43	293	0.04802	10.584	32819.6
44	293	0.04685	10.590	32860.1
45	293	0.04453	10.603	32939.7
46	293	0.0462	10.594	32882.3
47	293	0.04289	10.612	32996.4
48	293	0.0426	10.614	33006.6
49	293	0.04548	10.598	32907.1
50	293	0.04367	10.608	32969.6

Naive Bayes Model

Confusion Matrix

Training

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	390	16
>8	14	24

Validation

Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	258	13
>8	16	9

Overall

Column	Main Effect	Total Effect
Reason for absence 2 2	0.504	0.82
Pet	0.044	0.17
Disciplinary failure	0.046	0.151
Son	0.029	0.112
Month of absence	0.032	0.091
Social drinker	0.018	0.05
Work load Average/day Binned	0.016	0.039
Day of the week	0.014	0.036
Education	0.007	0.016
Seasons	0.004	0.008
Social smoker	0.001	0.003

Ensemble Model

Contingency Analysis of Absenteeism time in hours Binned 2 By Ensemble Model Validation=Validation

Mosaic Plot

Contingency Table

Absenteeism time in hours Binned 2

Ensemble Model	Count / Total % / Col % / Row %	0—8	>8	Total
>8	Count	3	7	10
	Total %	1.01	2.36	3.38
	Col %	1.09	33.33	
	Row %	30.00	70.00	
0—8	Count	272	14	286
	Total %	91.89	4.73	96.62
	Col %	98.91	66.67	
	Row %	95.10	4.90	
Total		275	21	296
		92.91	7.09	

Contingency Analysis of Absenteeism time in hours Binned 2 By Logistic Regression Classification Validation=Validation

Mosaic Plot

Contingency Table

Absenteeism time in hours Binned 2

Logistic Regression Classification	Count / Total % / Col % / Row %	0—8	>8	Total
>8	Count	8	11	19
	Total %	2.70	3.72	6.42
	Col %	2.91	52.38	
	Row %	42.11	57.89	
0—8	Count	267	10	277
	Total %	90.20	3.38	93.58
	Col %	97.09	47.62	
	Row %	96.39	3.61	
Total		275	21	296
		92.91	7.09	

Contingency Analysis of Absenteeism time in hours Binned 2 By Naive Bayes Predictions Validation=Validation

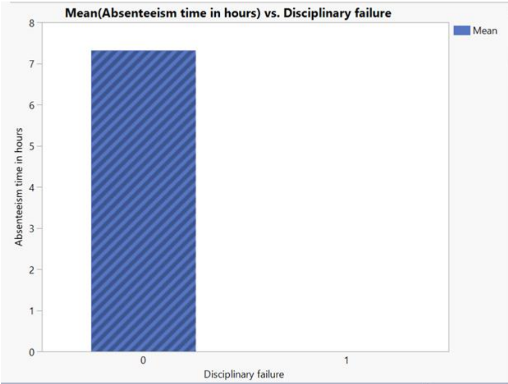
Mosaic Plot

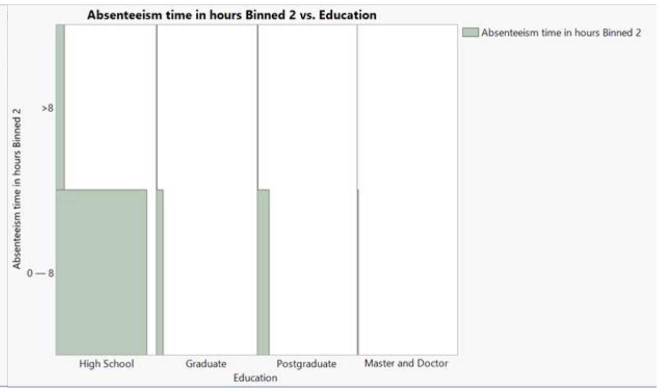
Contingency Table

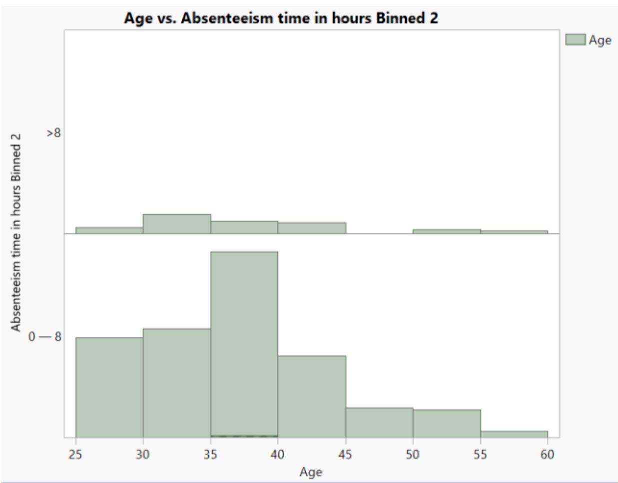
Absenteeism time in hours Binned 2

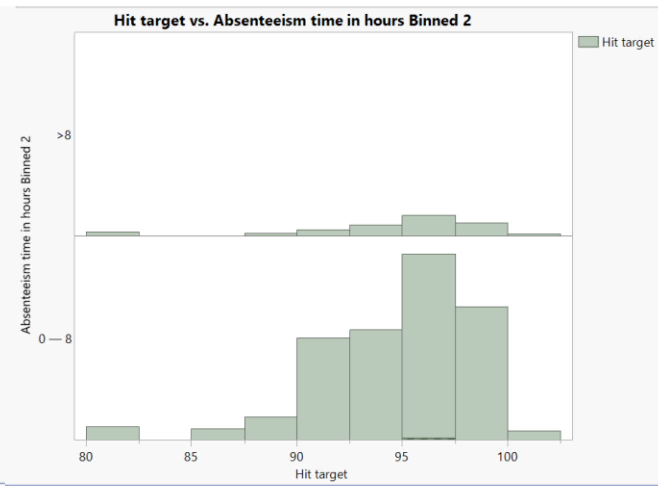
Naive Bayes Predictions	Count / Total % / Col % / Row %	0—8	>8	Total
>8	Count	12	13	25
	Total %	4.05	4.39	8.45
	Col %	4.36	61.90	
	Row %	48.00	52.00	
0—8	Count	263	8	271
	Total %	88.85	2.70	91.55
	Col %	95.64	38.10	
	Row %	97.05	2.95	
Total		275	21	296
		92.91	7.09	

Model Insights


Mean(Absenteeism time in hours) vs. Disciplinary failure
Mean
Absenteeism time in hours
Disciplinary failure
0
1
2
3
4
5
6
7
8


Absenteeism time in hours Binned 2 vs. Education
Absenteeism time in hours Binned 2
>8
0 — 8
High School
Graduate
Postgraduate
Master and Doctor
Education


Age vs. Absenteeism time in hours Binned 2
Age
Absenteeism time in hours Binned 2
>8
0 — 8
25
30
35
40
45
50
55
60


Hit target vs. Absenteeism time in hours Binned 2
Hit target
Absenteeism time in hours Binned 2
>8
0 — 8
80
85
90
95
100