Absenteeism at Work

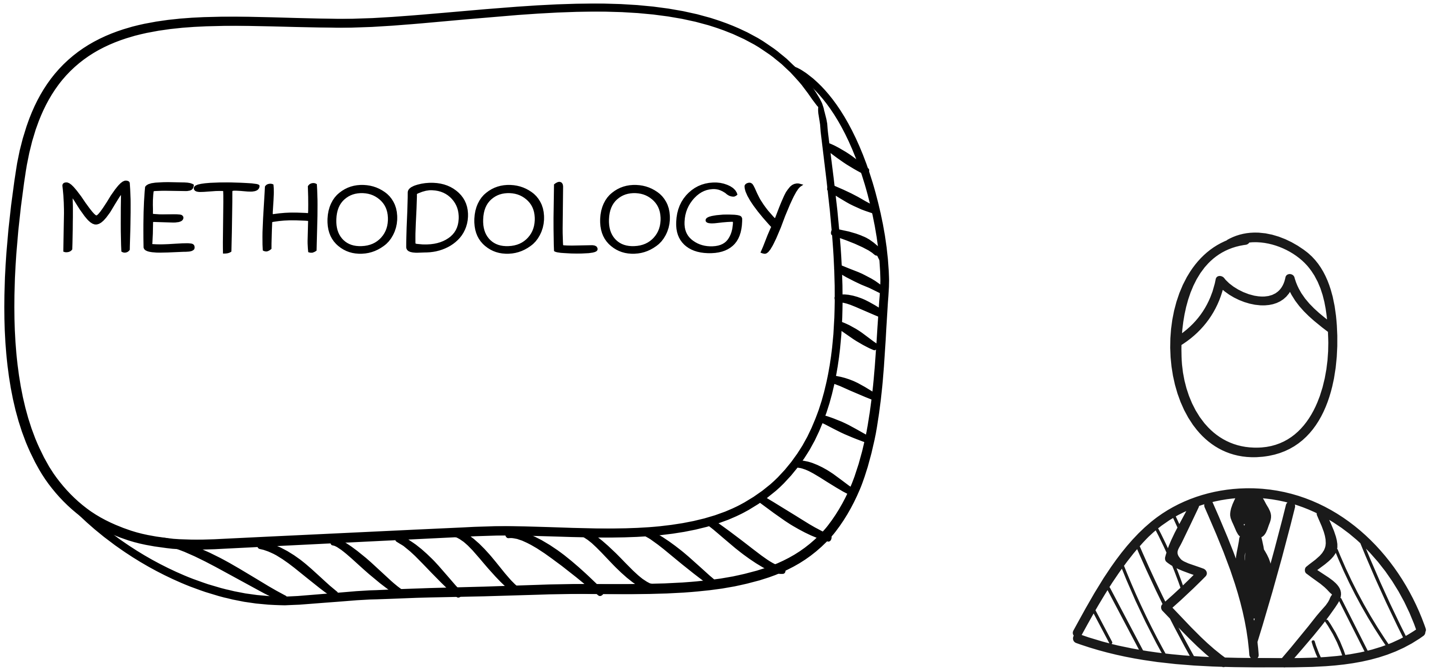
OPIM 5604 - Predictive Modeling

Team #3

TABLE OF CONTENTS

Problem Statement

- **What factors influence the duration of a single absenteeism event?**
- **Absenteeism is unavoidable**
    - Costly $$$
    - Number of packages delivered and delivery times
    - Affects other employees
    - Cycle of absenteeism and decreased productivity


METHODOLOGY

01 Sample

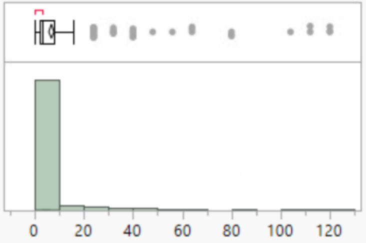
- **Absenteeism at Work Dataset:**
  - UCI Machine Learning Repository's archive of clean tabular datasets
  - Collected from a courier service in Brazil by PHD students at the Universidade Nove de Julho
  - Contains 21 columns and 740 rows describing instances of absenteeism, general information about employees, and reasons for absenteeism recorded over the course of 3 years
  - Bias: Dataset contains info from 38 unique employee IDs. We decided to treat each absentee event as unique, rather than focus on unique employees.

02 Explore

- Target Variable: Absenteeism time in Hours
  - ***Distinction between 0 to 8 hours and >8 hours***
- Predictive Variables
  - ***Most Significant:***
    - *‘Reason for Absence’ 17%*
    - *‘Height’ 14%*
    - *‘Day of the Week’ 12%*
    - *‘Disciplinary Failure’ 12%*
    - *‘Son’ 11%.*

Distributions

Absenteeism time in hours



Quantiles

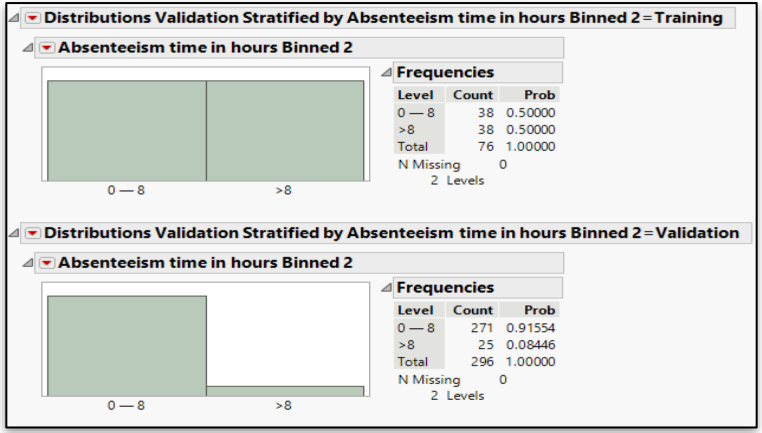
100.0%	maximum	120
99.5%		114.36
97.5%		40
90.0%		8
75.0%	quartile	8
50.0%	median	3
25.0%	quartile	2
10.0%		1
2.5%		0
0.5%		0
0.0%	minimum	0

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

03 Modify

- **Target Variable**
  - Absenteeism Time in Hours:
    - Binned 0 to 8 hours and > 8 hours
    - Reclassified as nominal variable
- **Partition**
  - 60% Training /40% Validation
  - Extra Validation Column
- **Predictive Variables**
  - Outliers - Age and Height
  - Redundant Variables - BMI
  - Workload Average/Day (40,000 bins) - 38 unique values to 5 unique ranges.



04 Model & Results

7 Predictive Models

Regression

Decision Tree

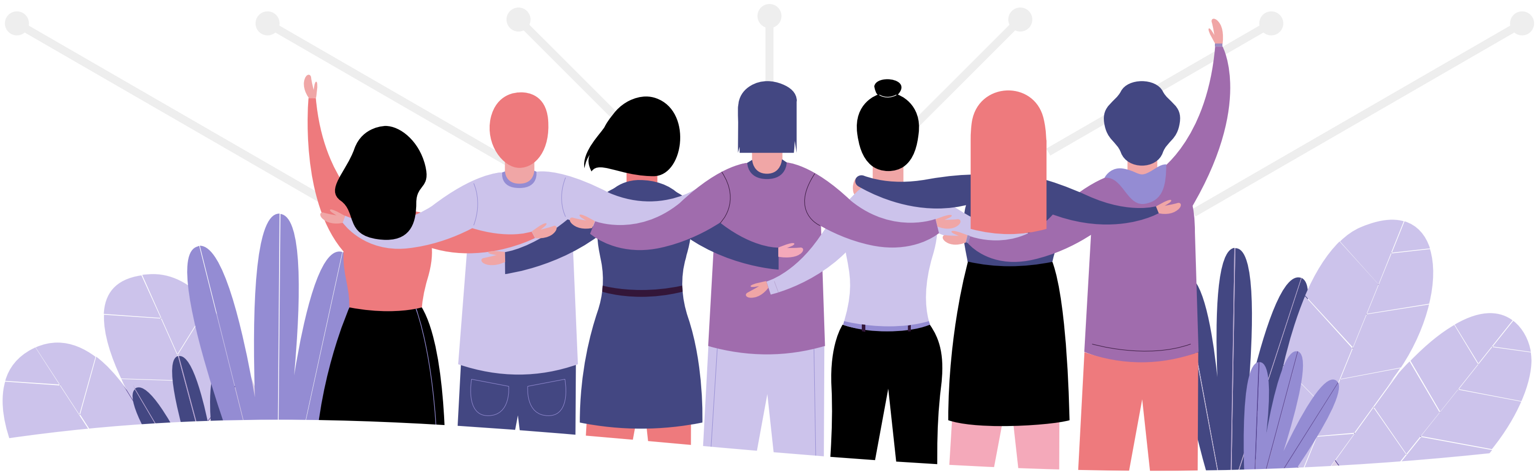
Neural Network

Discriminant Analysis

KNN

Naive Bayes

Ensemble



Regression Models

Linear Regression:

- Predictor Variables - All
- Initially poor metrics
- Using PCA, pared down under-performing variables for a better fit

Logistic Regression:

- Predictor Variables - All
- Immediately promising
- Overfitting: training misclassification rate of .045, validation misclassification rate of .1092
- Primary components flattened our misclassification rates across our tests

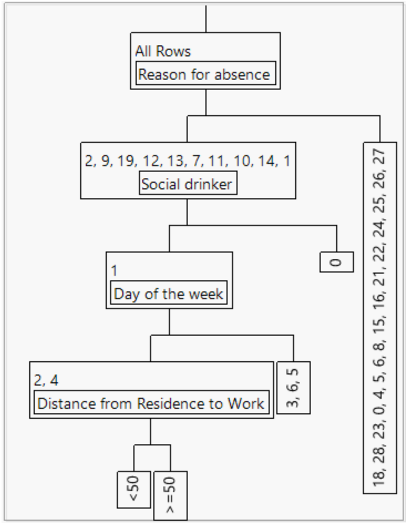
Nominal Logistic Fit for Absenteeism time in hours Binned 2

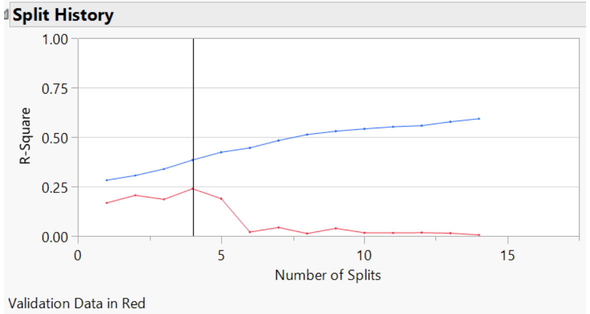
Effect Summary

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

- **Predictor Variables - All**
- Optimum Tree with 4 splits
- Significant predictors: Reason for absence, Social Drinker, Day of the Week and Distance from residence to work.





Confusion Matrix

Training

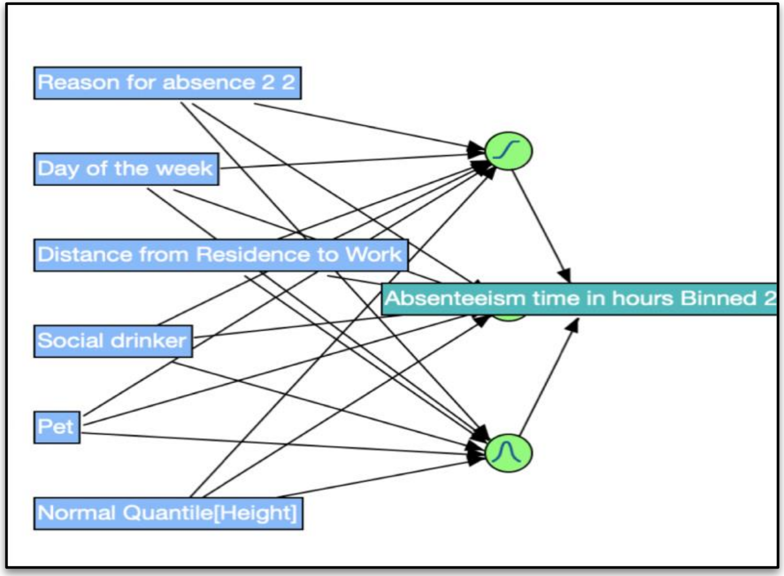
Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	Predicted Count >8
0 — 8	395	7
>8	23	19

Validation

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

Neural Network Model

- Predictor Variables:
  - 'Reason for absence', 'day of the week', 'distance from residence', 'social drinker', 'pet', and 'height'
- Hidden layer with 3 nodes:
  - 1 Tanh
  - 1 Linear
  - 1 Gaussian



Confusion Matrix

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

Confusion Rates

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

Confusion Rates

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

Discriminant Analysis Model

- **Predictive Variables - All**
- The model type is not a good fit for our data
  - Few normally distributed predictors
  - Unequal correlation to the target variable
- Better results from adjusting cutoff

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

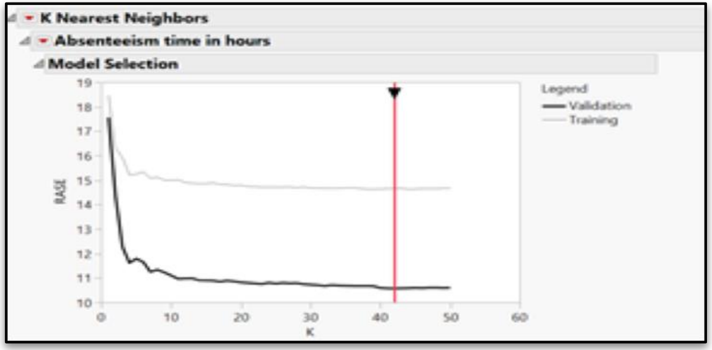
Validation

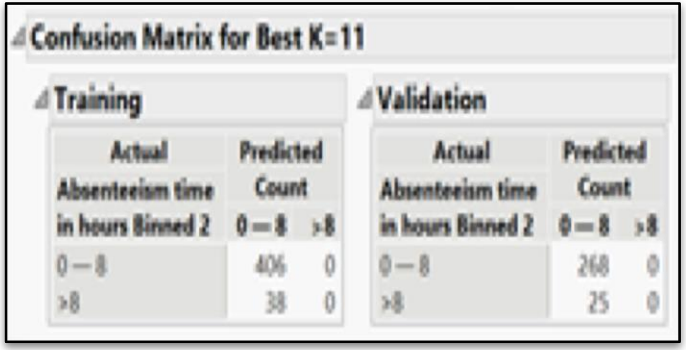
Actual Absenteeism time in hours Binned 2	Predicted Count 0 — 8	>8
0 — 8	224	47
>8	14	11

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

KNN Model

- Continuous Target Variable
  - Predictor Variables - All
  - Overfitting to training data
- Categorical Target Variable
  - Predictor Variables - All
    - Misclassified all record of interest





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

Naive Bayes Model

- Predictor Variables:
  - All categorical variables
- Results:
  - Overfit to training data
  - Important predictors: Reasons for Absence, Pet, and Disciplinary Failure

Confusion Matrix

Training

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

Validation

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

Cols	Absenteeism 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

- Took a majority vote of the predicted classes by our 5 major models for ensemble classification

05 Assess

- **Assessment Criteria:**
  - Misclassification Rate
  - RMSE
  - RSquared
  - RASE

Results

Ensemble Model is our best performing model.

- Misclassification rate: 0.0473
- RSquare: 0.4010
- RMSE: 0.2089
- Overall accuracy: 95.27%
- Accuracy of “>8” hours of absenteeism predictions: 70%
- Accuracy of “0 - 8” hours of absenteeism predictions: 95%

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

- ‘Reason for Absence’, ‘Day of the Week’, ‘Disciplinary Failure’ and ‘Son’ had the most significant relationships with ‘Absenteeism in Hours.’
  - The most common ‘Reasons for Absence’ are (23) Blood Donation 20%, (28) Dental Consultation 15%, and (27) Physiotherapy 9%.
  - Employees are more likely to be absent on (2) Monday 21% than on (5) Friday 16%.
  - Those who never received disciplinary failure are more likely to be absent (95%) than those who haven’t (5%).
- Employees who are between the age of 30 to 45 years are late to work more often.
- Employees with a higher level of education are more likely to be present to work and on time.
- 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.

Recommendations

Attendance Policies

- Defines each type of absence and address attendance tracking
- Fair to both employees and employers

Orientation Programs

- Highlight the consequences of absenteeism

Reward System

- Enable and reward good behavior

Health Policies

- Provide coverage for illnesses prevalent among employees

Flexible Work Options

- Work from home for office workers

Employee Screening

- Screen employees for absenteeism and monitoring absenteeism

THANK YOU!