SAS Project: PA Vehicle Accidents Report

Matt Chylack, Ben Phillippy, Jayden Carlucci, and Matt Granato

Executive Summary

- Analyze the variables that go into car crashes in pennsylvania from the years 2016-2021
- Define/categorize variables which have most influence on severity of car crash
- Help state/local officials minimize severity of crashes

Why It Matters

- Make people more aware of crash risk in order to help decrease the amount of crashes
- Help people become more aware of dangerous road conditions
- Influence drivers to be safer/drive in adequate conditions
- Influence lawmakers to make appropriate driver safety laws to decrease injury/deaths tolls

Project Motivation

- Understanding what causes long traffic backups in Pennsylvania
- Help drivers understand what variables cause the most risk for accidents
- Give local organizations reasons to make appropriate changes to roadways/appropriate warnings for inclement weather

Data Description

- Original dated consisted of info from 49 states in 2.7 million records
- Multiple methods used to collect data of incident-Traffic cameras, police reports, traffic sensors
- All data was reported to Department of Transportation

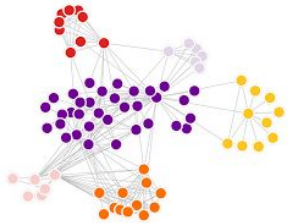
Data Cleanings

- Report consists of data solely from Pennsylvania
- 41,850 rows/46 columns
- Omitted non-nullable values

Data Preparation Activities

- Imported file into Sas
- Edited variables to appropriate Role and Level
- Assigned Distance_mi as target variable; Distance (miles) of traffic backup during accident
- Rejected Severity-Will be known afterwards by distance of backup

Models/Diagrams Used



- First added node was Cluster Analysis
- Found that more significant accidents occur during daytime
- Accidents more likely to occur near sunrise/sunset
- More likely to have significant accident in foggy condition or thunderstorm
- Distance_mi and true/false Sunrise/Sunset are top 2 variables of significance

Variables of Highest Significance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Distance_mi_		2	8	1.00000
Sunrise_Sunset	Sunrise_Sunset	0	2	0.86592
Pressure_in_		2	4	0.73157
Civil_Twilight	Civil_Twilight	1	0	0.70890
Temperature_F_		3	2	0.68775
Astronomical_Twilight	Astronomical_Twilight	1	3	0.68010
Visibility_mi_		5	3	0.62379
Weather_Condition	Weather_Condition	1	7	0.62299
Give_Way	Give_Way	1	0	0.17916
Precipitation_in_		4	2	0.15029

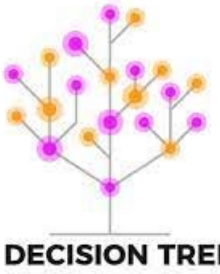
Frequency of Cluster ▼	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Distance_mi_	Precipitation_in_	Pressure_in_	Temperature_F_	Visibility_mi_
10635	0.168607	17.51297	6	1.772007	0.455364	.0004731	29.67635	76.21837	9.80038
9958	0.293106	14.97431	6	3.485089	0.755494	.0007814	29.44101	50.39643	9.578563
7790	0.183498	12.66587	1	1.772007	0.460964	.0004601	29.76023	45.61048	9.752142
4733	0.237715	14.56003	1	2.21747	0.633576	0.000688	28.4535	60.62519	9.668727
2977	0.2801	13.08783	1	2.885551	3.300992	0.000572	29.53975	63.59457	9.67664
2297	0.580904	30.34441	6	4.313358	0.864281	0.033475	29.27987	54.40227	3.851199
460	0.415146	30.31451	6	9.401933	0.801993	0.002095	29.44022	56.74573	9.281938
157	.	0.432349	4	28.83545	0.068688	1.26E-18	30.14	75	3
66	.	3.962106	4	29.2965	0.254652	0.113333	29.63231	60.38	5.4
44	.	3.469149	4	28.77976	1.552159	0.005	29.92615	34.15	3.45
40	.	5.410877	3	20.4234	3.84905	1.48E-19	29.87875	20.68571	8.571429
23	.	2.85324	4	28.84492	0.615348	1.48E-19	28.22	58.5	3
23	.	0.706679	3	28.77957	2.29687	1.48E-19	29.095	64	5

Tests Ran

- Ran three tests-Neural Network, Decision Tree, & Regression
- Used to predict distance of backup
- Used Model comparison-found Decision Tree to be most accurate of predicting crash backup distance
- Had lowest AVG squared error of 2.37

Testing Various Decision Trees

- Adjusted max number of branches & maximum depth
- Best model came out to be one having 3 branches and depth of 6
- Had average Squared Error of 2.37



Findings

- Variables within our dataset made it difficult to accurately predict factors that cause traffic delays for accidents
- Concluded that more traffic on roads results in higher likelihood of accident delaying traffic
- Majority of these large traffic delays occurred in daytime hours followed by near sunrise/sunset