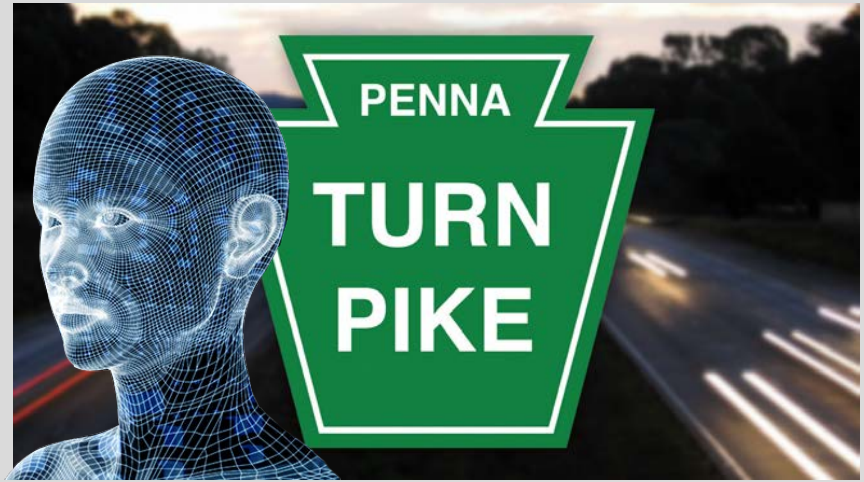
ACCIDENT RISK PREDICTION



Proof of Concept

for

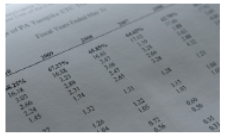
Pennsylvania Turnpike Commission
700 South Eisenhower Blvd.
Middletown, PA 17057

PTC INFORMATION TECHNOLOGY

Background

BA



DRIVING TOWARD SUCCESS

STRATEGIC DRIVER: FINANCIAL

Goal: Maintain a sound financial position.

Our staff of financial professionals is deeply committed to managing our customers' toll dollars and the Commission's overall financial position so that we continue to provide a safe and valued travel experience every day while making wise and needed investments in our system's infrastructure. We recognize that the national and global economies, as well as local and state level policy decisions, can have a direct impact on our financial position and we have a responsibility to either help guide or respond accordingly to these external forces.

Objectives and Measures:

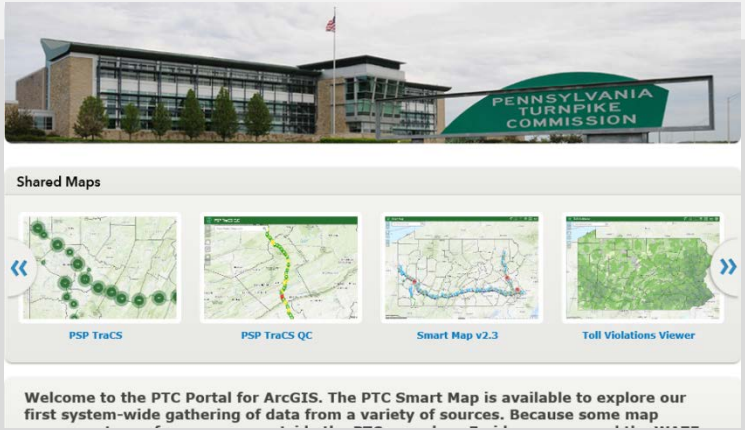
1. **Maintain and provide support to our credit rating.**
   Measures: Bond rating; Debt services coverage ratio
2. **Improve operational efficiencies.**
   Measures: Percentage of growth in the operating budget; Overall variance between approved operating budget and the year-end budget spending
3. **Maintain or increase toll revenue amount.**
   Measures: Toll revenue amount; Capture rate = Net Toll Captured (Net Toll Revenue + Fee Revenue - Bad Debt and Revenue Adjustments) and Percent of Net Tolls Captured (Net Tolls + Fees - Bad Debt and Revenue Adjustments/Net Tolls + Fee Revenue); Non-toll revenue amount
4. **Promote and analyze the Turnpike's ability to implement innovations.**
   Measure: Number of innovations and innovative practices evaluated by the Innovation Council, AIM, or other PA Turnpike committee

BUSINESS PROBLEM

pennsylvania STATE POLICE

CADS

INRIX



100s of Variables

10 Years of Data
700,000 Accidents
1,200+ Segments

Impossible to Manually Analyze

Train a Machine to do it?

PARADIGMS OF ANALYTICS

Artificial Intelligence

Business Intelligence, Big Data..

Business Question	**What happened?** →	**What is Happening?** →	**Why is it happening?** →	**What will Happen?** →	**What action to take?**
Tools	Operational Reports Pre-defined Queries KPIs Scorecard	Real-Time Dashboards Automated Alerts	Statistical Models Root Cause Analysis	Predictive Analytics Machine Learning Deep Learning	Optimization Reinforcement Learning Decision Automation
Examples	*Post-Mortem Accident Investigation*	*An Accident just happened along the Turnpike*	*What are the causes of the accidents at a particular Milepost?*	*Conditions Indicate that Milepost segment T 069.8 EB has potential for an accident*	*Place emergency personnel at probable locations to reduce response time by 25%*

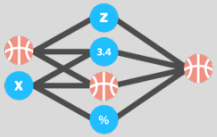
TYPES OF ARTIFICIAL INTELLIGENCE

Reasoning

Generating conclusions from available knowledge using logical techniques such as deduction and induction

E.g. Scheduling

Machine Learning



Ability of systems to Learn from historical data & Predict outcomes without being explicitly programmed. Experience increases by time

E.g. Fraud Detection

Knowledge Representation

Representing information about the world in a form that a computer system can utilize to solve complex tasks

E.g. Semantic Nets

Perception

Ability to perceive the world through input from sensors: Cameras, Microphones, Images, Sonar & more

E.g. Facial Recognition

Robotics

Ability of Robots to accurately perceive their location & surroundings, plan their path, & manipulate objects

E.g. Robots

Natural Language Processing

Understanding Natural Language & extracting entities, sentiment, & emotions

E.g. Smart Agents

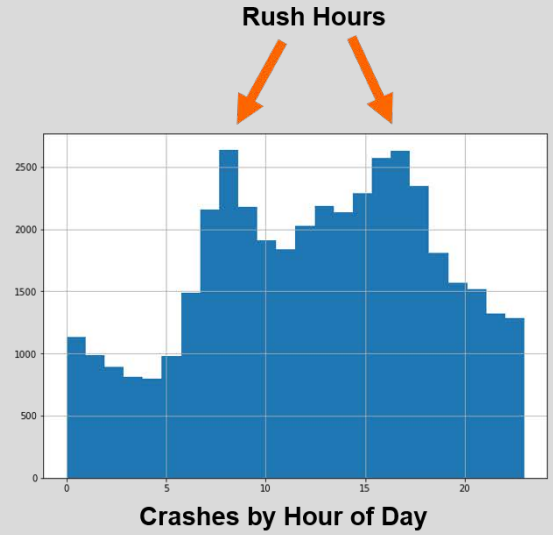
Machine Learning Steps

STEP 1: EXPLORE THE DATA

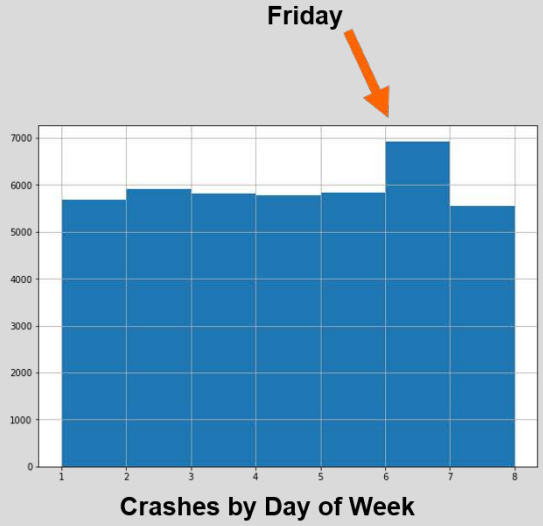
Use GIS tools to Identify patterns to determine a strategy

EXPLORING THE FEATURES

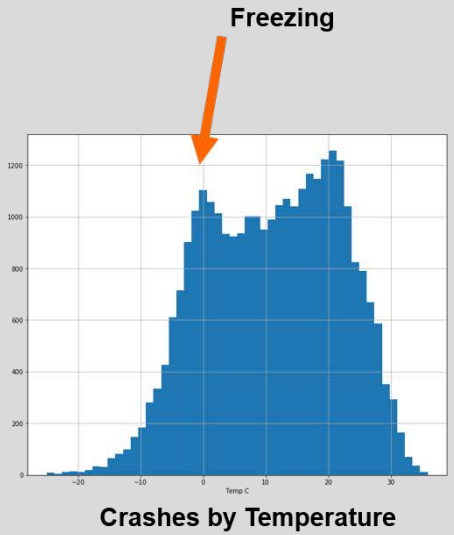
- Some patterns in features can be used to identify times and places with high risk



Crashes by Hour of Day



Crashes by Day of Week



Crashes by Temperature

STEP 2: DATA PREPARATION

Temperature
Sun, Mon, Fri..

Wind Speed
Fast, Slow..

Visibility
High/Low

Snow Depth
High/Low

Day of the Week
Sun, Mon, Fri..

Time of the Day
12:45, 23:00

Month
Feb, Dec..

Road Width
20-30 M

Road Alignment
Straight / Curved

Proximity to
Intersections

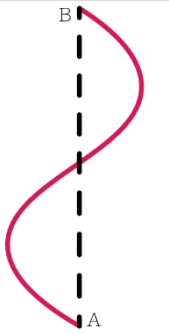
Speed Limit
120 km/h

Sun Direction
East, West

Daily Traffic
AADT

Proximity to
Billboards

DEFINITIONS



Sinuosity - length of straight line connecting endpoints divided by segment length

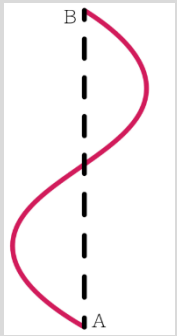
Source: https://medium.com/a-r-g-o/rad-roads-using-osmnx-68fc8b15f046

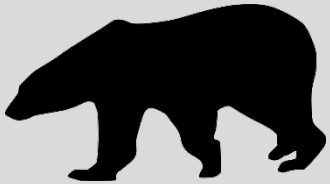
Orientation - angle of the line connecting the endpoints respecting direction of travel

GIS DATA PREPARATION

Roads

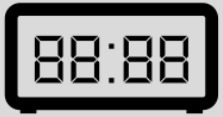
= 1 mile segments + Population per segment



+  +  +  + INRIX

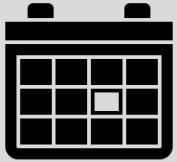
ACCIDENT DATA PREPARATION

Time Features



Hour of the Day

Day of Week



Day of Year

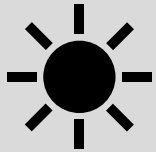
Road Segment & Accident

Distance Features

Distance to Bridge

Distance to Interchange

Sun Features



Position of Sun

Solar Angle relative to road

STEP 3: FEATURE ENGINEERING

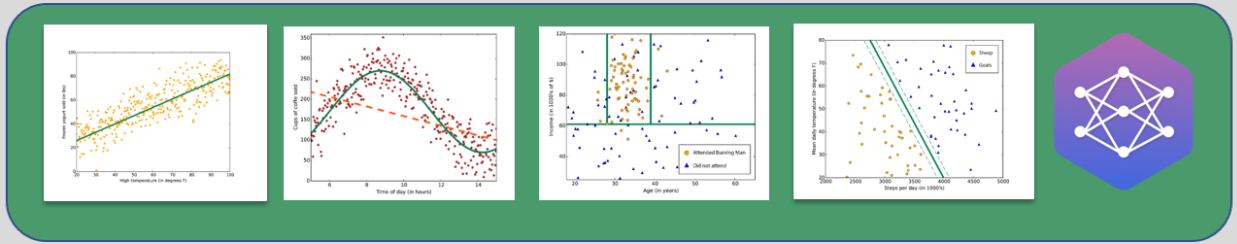
Features | **Output (label)**

Training Data (Historical)

For **Learning..**

Segment Type	Proximity to Interchange	Time of Day	Weather	Incident
Highway	0.1 M	Morning	Raining	**Car Accident**
Tunnel	0.3 M	Evening	Sunny	**None**
Bridge	0.2 M	Noon	Foggy	**Car Accident**

Learning Models



Trained Model

FEATURES TO USE IN MODELING

Static

Properties that do not change over time (road curves, number of accidents, etc.)

Dynamic

Properties that change with time (sun angle, weather, traffic, etc.)

PTC STATIC FEATURES USED

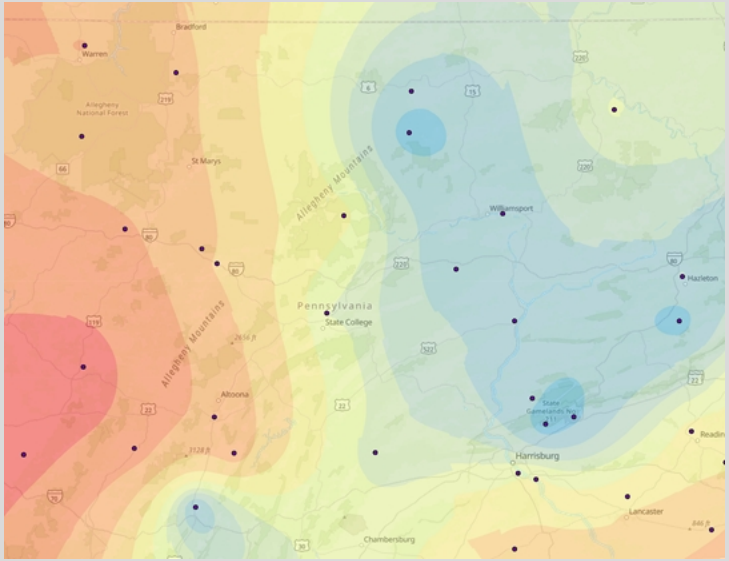
Feature	Type	Source
Road Sinuosity	Static	PTC Geodatabase
Road Orientation	Static	PTC Geodatabase
Historical Accidents	Static	PA State Police TrACS
Bridges	Static	PTC Geodatabase
Tunnels	Static	PTC Geodatabase
Population Density	Static	Esri Living Atlas
Median Age	Static	Esri Living Atlas
Distance to Bridges	Static	Derived
Distance to Tunnels	Static	Derived
Distance to Interchanges	Static	Derived
Distance to Ramps	Static	Derived
Distance to Toll	Static	Derived
Distance to DMS	Static	Derived
Distance to Service Plaza	Static	Derived
Bear Encounters	Static	Computer Aided Dispatch
Deer Kills	Static	Computer Aided Dispatch

PTC DYNAMIC FEATURES USED

Feature	Type	Source
Air Temperature	Dynamic	MesoWest
Precipitation	Dynamic	MesoWest
Wind Gust	Dynamic	MesoWest
Dew Point Temp	Dynamic	MesoWest
Wind Speed	Dynamic	MesoWest
Relative Humidity	Dynamic	MesoWest
Day of Year	Dynamic	Attribute (where available)
Day of Week	Dynamic	Attribute (where available)
Time of Day	Dynamic	Attribute (where available)
Solar Elevation	Dynamic	Python Library
Solar Azimuth	Dynamic	Python Library
Solar Angle Relative to Road Orientation	Dynamic	Derived
Speed, Delta, Reference, Score, Congestion, # of Lanes	Real-Time	INRIX

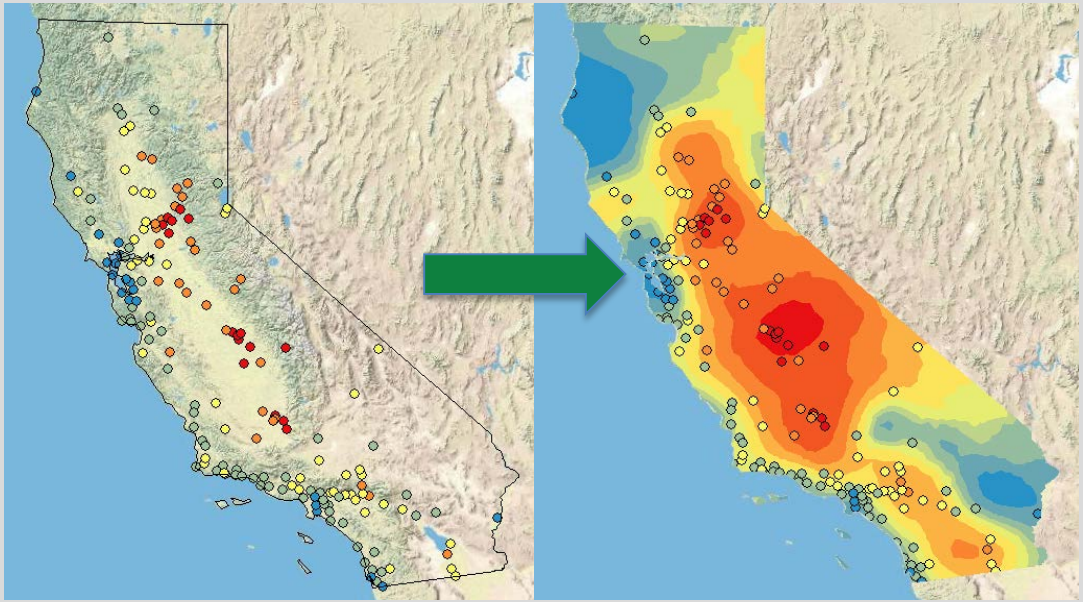
WEATHER FEATURES

- Obtaining historic weather was a challenge
- RWIS not available
- Inconsistent data sources
- Weather stations not on road



WEATHER APPROACH

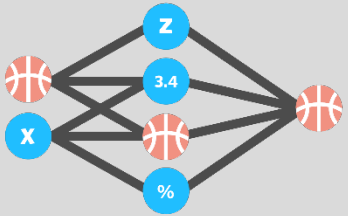
Interpolation – the closer a weather station to incident, the more similar the conditions



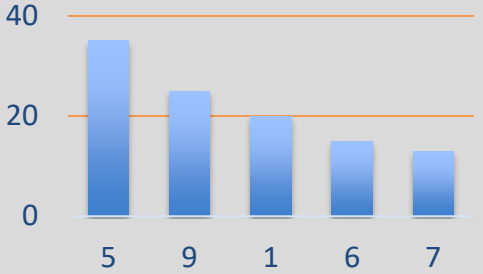
STEP 4: MODEL CREATION

Experiment with different models to find the best results

Raw Data

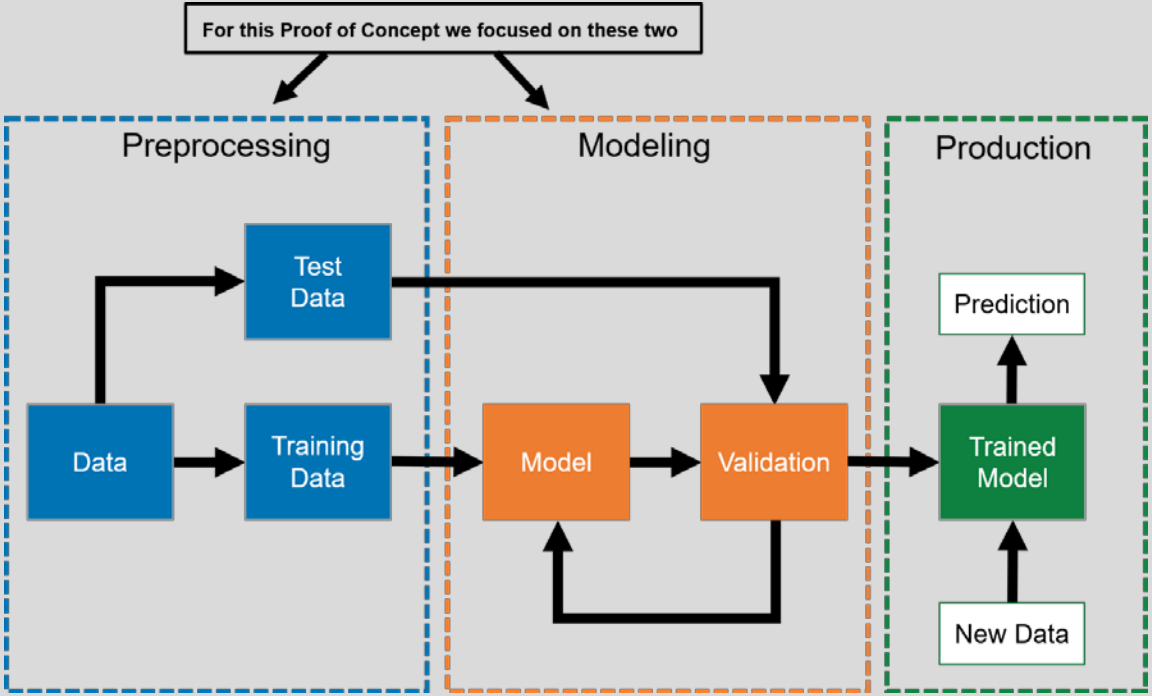


Model



Top Correlating Features

MACHINE LEARNING WORKFLOW



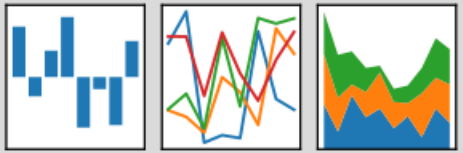
TECHNICAL APPROACH

ArcGIS Pro

ArcGIS API for Python
ArcPy

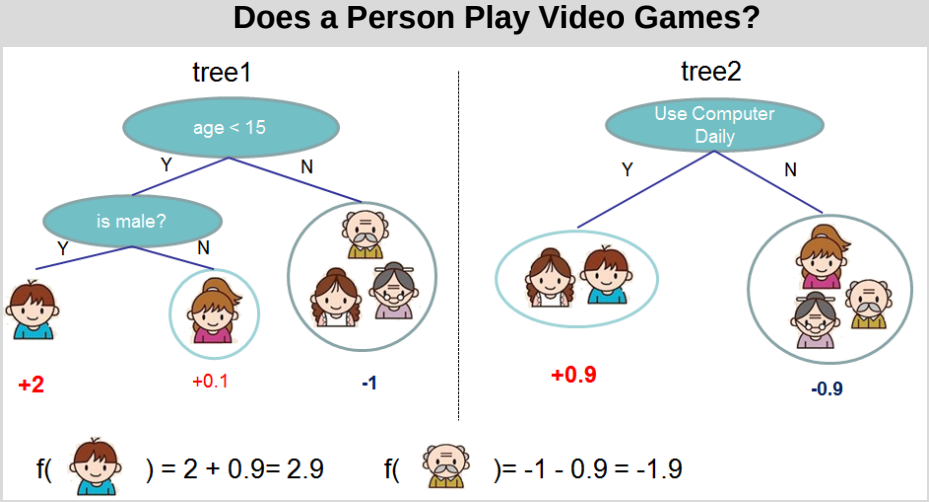
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



XGBOOST

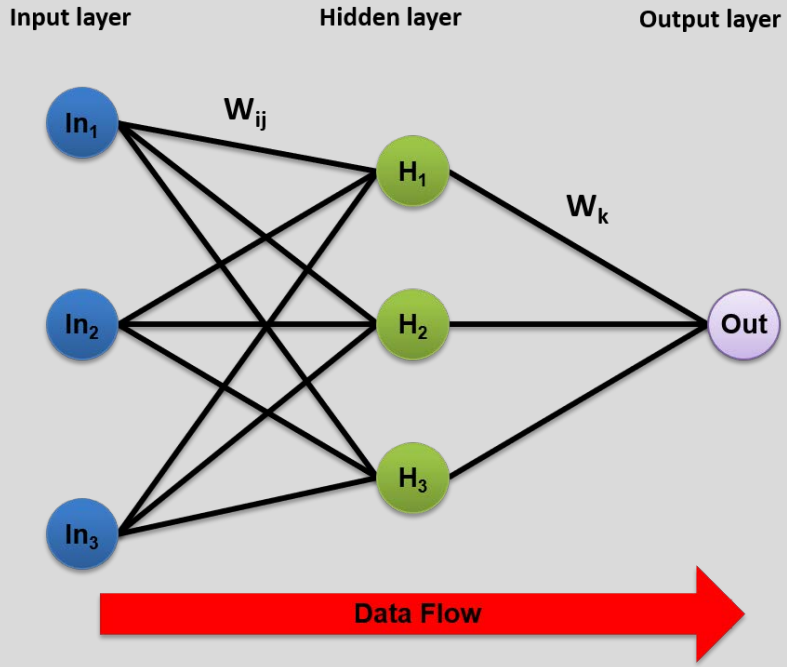
- Algorithm *learns* to build decision trees.
- After many trees have been built, predictions are made by combining the results of the entire ensemble.
- Combines many *Weak Learners* in to one *Strong Learner*.



Xgboost.readthedocs.io

NEURAL NETWORKS

- Roughly inspired by the Human brain.
- Constructed from a series of *nodes* and *connections* (or *weights*) arranged in *layers*.
- Data goes in the input nodes and the W's (the *weights*) determine how strongly connected different nodes are.
- After the network learns the W's, the output can be used for prediction.
- In practice networks can have many layers (Deep Networks) and hundreds of nodes per layer.



NAÏVE BAYES

- Treat each feature as a “probability”
- Combine all features to get a measure of “risk”

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Source: https://www.saedsayad.com/naive_bayesian.htm

MODEL EVALUATION

Model	Pros	Cons
XGBoost	• Ranks features by importance, • Few parameters to adjust • Low overhead to get running (less code, less data prep)	• May not pick up on very complicated relationships, • Tendency to “overfit” the data
Neural Network	• Powerful model that can find very complex relationships	• Many parameters to adjust, must prepare data carefully • Blackbox (difficult to understand what is going on inside)
Naïve Bayes	• Very simple model to implement and to understand • Easy to adjust	• May not find complex relationships • Must have completely uncorrelated features (e.g. may have to use less features than other models)

Using the Model

RISK OVER TIME VISUALIZATION

WEATHER EVENT

Preparing for Machine Learning

STRUCTURED AND POPULATED DATA

Incident_Number	Crash_Date	Crash_Time	County_Code	Principal_Road_Speed_Limit
A01-1663764	1/18/2010	11	64	5
A01-1858102	10/22/2013	0	64	65
B01-1684810	12/13/2008	22	62	55
B02-1739963	10/25/2009	1	26	55
B02-1952337	2/5/2013	6	26	65
B02-1996781	10/21/2013	14	26	65
B02-2109435	5/8/2015	7	62	55
B02-2120467	6/30/2015	6	26	55
B03-1775353	5/7/2010	0	2	65
B03-1854590	7/4/2011	22	2	55
B03-1867452	9/12/2011	6	2	55
B03-1870321	9/28/2011	3	62	65
B03-1896689	2/29/2012	8	2	55
B03-1972068	5/31/2013	14	2	55
B03-1998951	11/2/2013	23	2	0
B03-2039952	5/19/2014	20	62	55
B03-2042820	6/2/2014	8	2	55
B03-2046385	6/17/2014	6	2	55
B03-2058339	8/13/2014	8	2	65
B03-2064803	9/18/2014	5	2	55
B03-2079054	12/2/2014	9	2	55

IDEAL STRUCTURE

INCIDENT NUMBER	Roadway	Month	Year	INCIDENT TIME	Hour	DAY OF WEEK	WEATHER CONDITIONS	RD SURFACE COND
T09-9049967	TURNPIKE MAINLINE	12	2015	0.06875	1	THURSDAY	NO ADVERSE CONDITIONS	DRY
T09-9048353	TURNPIKE MAINLINE	4	2015	0.272916667	6	MONDAY	NO ADVERSE CONDITIONS	DRY
T09-9048079	TURNPIKE MAINLINE	2	2015	0.753472222	18	MONDAY	SNOW	WET
PA 2017-847393	TURNPIKE MAINLINE	8	2017	0.402777778	9	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9049046	TURNPIKE MAINLINE	7	2015	0.689583333	16	SUNDAY	NO ADVERSE CONDITIONS	DRY
T09-9050849	TURNPIKE MAINLINE	6	2016	0.267361111	6	WEDNESDAY	RAIN	WET
T09-9049356	TURNPIKE MAINLINE	9	2015	0.422222222	10	TUESDAY	NO ADVERSE CONDITIONS	DRY
T09-9051061	TURNPIKE MAINLINE	7	2016	0.666666667	16	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-904091	TURNPIKE MAINLINE	8	2017	0.479166667	11	THURSDAY	NO ADVERSE CONDITIONS	DRY
T09-9048523	TURNPIKE MAINLINE	5	2015	0.40625	9	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9048346	TURNPIKE MAINLINE	4	2015	0.847222222	20	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-836621	TURNPIKE MAINLINE	8	2017	0.339583333	8	SUNDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-831596	TURNPIKE MAINLINE	8	2017	0.777083333	18	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-839431	TURNPIKE MAINLINE	8	2017	0.348611111	8	MONDAY	RAIN	WET
PA 2017-579594	TURNPIKE MAINLINE	6	2017	0.1125	2	THURSDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-68382	TURNPIKE MAINLINE	1	2017	0.555555556	13	SUNDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-6206	TURNPIKE MAINLINE	1	2017	0.458333333	11	MONDAY	NO ADVERSE CONDITIONS	WET
PA 2017-839581	TURNPIKE MAINLINE	8	2017	0.360416667	8	MONDAY	RAIN	WET
T09-9050212	TURNPIKE MAINLINE	2	2016	0.1875	4	SATURDAY	NO ADVERSE CONDITIONS	DRY
T09-9050162	TURNPIKE MAINLINE	1	2016	0.482638889	11	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9048574	TURNPIKE MAINLINE	5	2015	0.621527778	14	WEDNESDAY	NO ADVERSE CONDITIONS	DRY

AVOID MANUAL DATA ENTRY

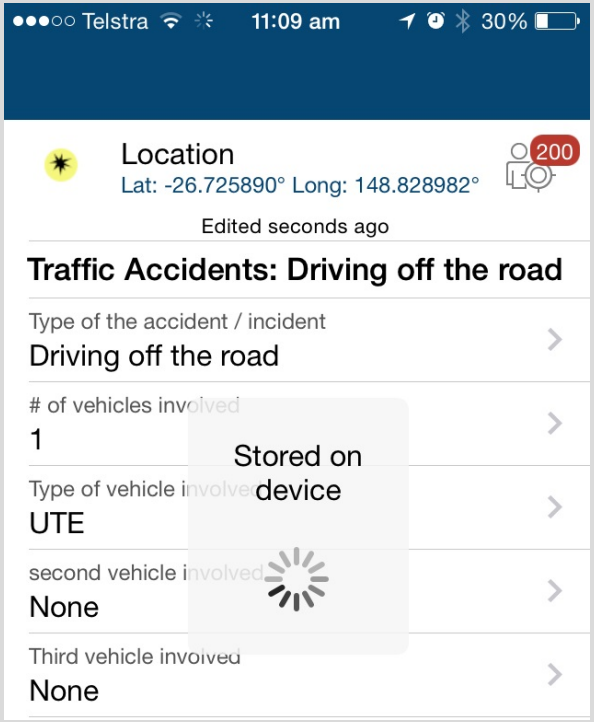
- Manual entry for join field

Console2 | AGENCY_EVENT

```
> select distinct(ptcpspcadnumber) from AGENCY_EVENT;
```

ptcpspcadnumber
PA20160000002
`
PA2016-52925
PA20163389
PA000000000
61850
2016-68795
69195
2016-72693
PA2016-75732
2016-75819
PA16-76465
PA1680446
PA81839
PA82051
83175
PA16-84984

GO DIGITAL IN THE FIELD



RECOMMENDATIONS

- Establish well-defined data model
- Create structured databases
- Enforce consistent data entry
- Capture well-structured data up front

PRESERVE GRANULAR DETAILS

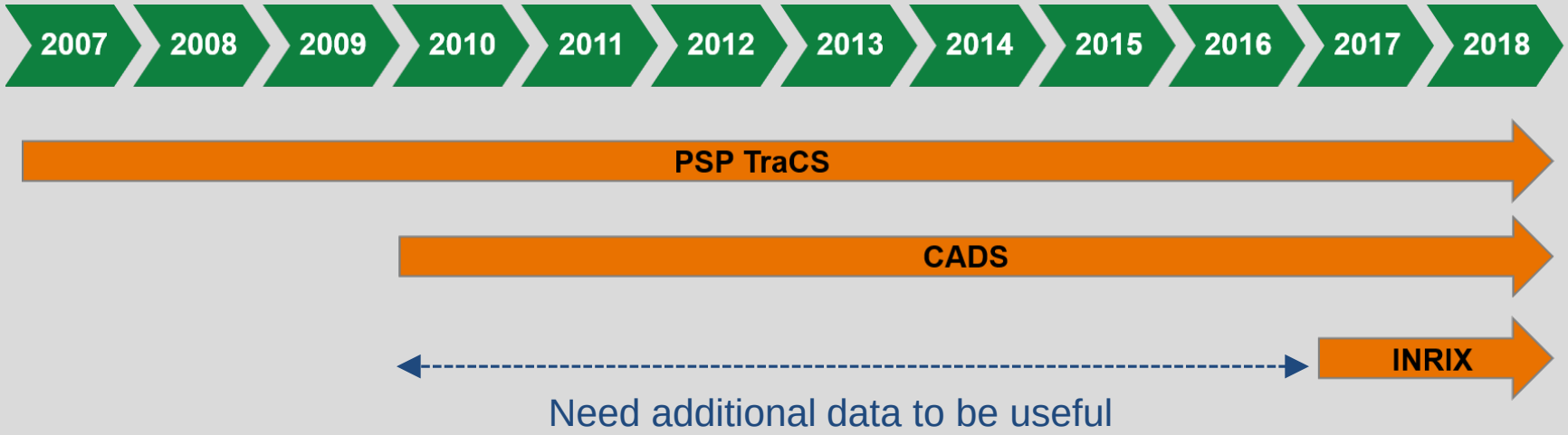
INCIDENT NUMBER	Roadway	Month	Year	INCIDENT TIME	Hour	DAY OF WEEK	WEATHER CONDITIONS	RD SURFACE COND
T09-9049967	TURNPIKE MAINLINE	12	2015	0.06875	1	THURSDAY	NO ADVERSE CONDITIONS	DRY
T09-9048353	TURNPIKE MAINLINE	4	2015	0.272916667	6	MONDAY	NO ADVERSE CONDITIONS	DRY
T09-9048079	TURNPIKE MAINLINE	2	2015	0.753472222	18	MONDAY	SNOW	WET
PA 2017-847393	TURNPIKE MAINLINE	8	2017	0.402777778	9	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9049046	TURNPIKE MAINLINE	7	2015	0.689583333	16	SUNDAY	NO ADVERSE CONDITIONS	DRY
T09-9050849	TURNPIKE MAINLINE	6	2016	0.267361111	6	WEDNESDAY	RAIN	WET
T09-9049356	TURNPIKE MAINLINE	9	2015	0.422222222	10	TUESDAY	NO ADVERSE CONDITIONS	DRY
T09-9051061	TURNPIKE MAINLINE	7	2016	0.666666667	16	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-904091	TURNPIKE MAINLINE	8	2017	0.479166667	11	THURSDAY	NO ADVERSE CONDITIONS	DRY
T09-9048523	TURNPIKE MAINLINE	5	2015	0.40625	9	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9048346	TURNPIKE MAINLINE	4	2015	0.847222222	20	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-836621	TURNPIKE MAINLINE	8	2017	0.339583333	8	SUNDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-831596	TURNPIKE MAINLINE	8	2017	0.777083333	18	FRIDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-839431	TURNPIKE MAINLINE	8	2017	0.348611111	8	MONDAY	RAIN	WET
PA 2017-579594	TURNPIKE MAINLINE	6	2017	0.1125	2	THURSDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-68382	TURNPIKE MAINLINE	1	2017	0.555555556	13	SUNDAY	NO ADVERSE CONDITIONS	DRY
PA 2017-6206	TURNPIKE MAINLINE	1	2017	0.458333333	11	MONDAY	NO ADVERSE CONDITIONS	WET
PA 2017-839581	TURNPIKE MAINLINE	8	2017	0.360416667	8	MONDAY	RAIN	WET
T09-9050212	TURNPIKE MAINLINE	2	2016	0.1875	4	SATURDAY	NO ADVERSE CONDITIONS	DRY
T09-9050162	TURNPIKE MAINLINE	1	2016	0.482638889	11	WEDNESDAY	NO ADVERSE CONDITIONS	DRY
T09-9048574	TURNPIKE MAINLINE	5	2015	0.621527778	14	WEDNESDAY	NO ADVERSE CONDITIONS	DRY

PRESERVE GRANULAR DETAILS

PENNSYLVANIA TURNPIKE COMMISSION
2017 MAINLINE AVERAGE ANNUAL DAILY TRAFFIC (AADT)
AND AUGUST AVERAGE DAILY TRAFFIC (ADT)
BETWEEN INTERCHANGES

Interchange	EB / NB AADT	WB / SB AADT	Total AADT	Percent Trucks	August EB / NB ADT	August WB / SB ADT	August ADT	August Percent Trucks
T 2 - T 30	12,272	N/A	12,272	23%	14,330	N/A	14,330	21%
T 30 - T 39	17,970	16,086	34,057	21%	20,723	18,624	39,347	21%
T 39 - T 48	19,853	17,915	37,768	20%	22,754	20,667	43,421	19%
T 48 - T 57	21,897	20,079	41,976	18%	24,788	22,936	47,724	18%
T 57 - T 67	24,752	22,998	47,751	17%	27,919	26,187	54,106	17%
T 67 - T 75	19,805	18,425	38,229	21%	22,706	21,431	44,137	20%
T 75 - T 91	19,665	18,221	37,886	29%	22,778	21,558	44,337	27%

BUILD UP DATA ARCHIVES

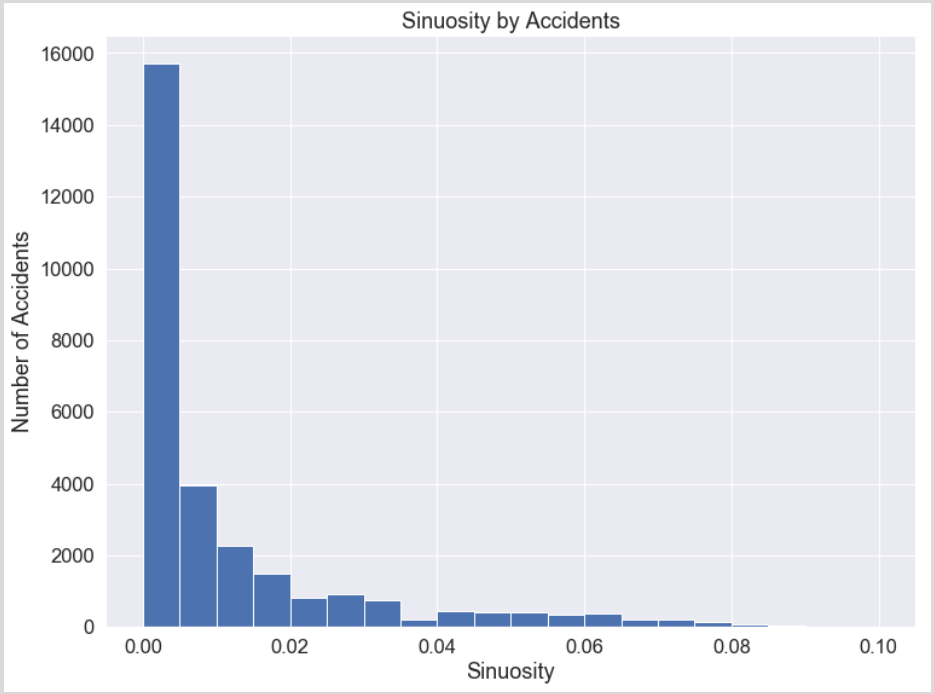


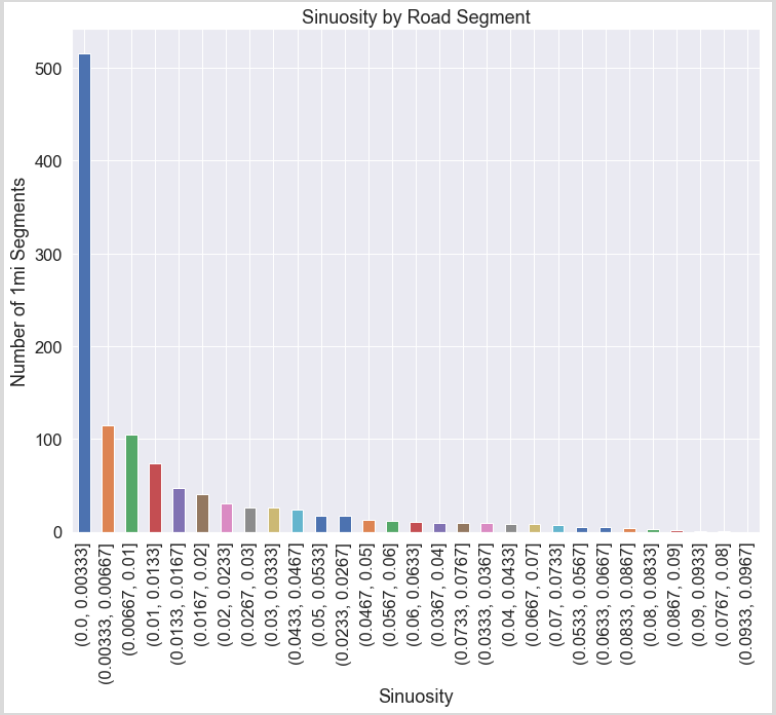
OTHER MEASURES

- Establish a feature pipeline that can be refreshed with new data
- Pre-join tables to avoid cross referencing
- Avoid missing data and inconsistent formatting (must be estimated or discarded from analysis)

Next Steps

MODEL IMPROVEMENTS - BIAS





HISTORICAL WEATHER

- Need reliable provider for “point in time” historical weather
  - Solar radiation
  - Precipitation rate
  - Precipitation depth
  - Precipitation type
  - Conditions (rainy, foggy,, etc.)
  - Temperature (Air, Dew Point, and ideally road surface)
  - Humidity
  - Windspeed/Direction
  - Visibility
  - Cloud Cover

EXPLORE CADS FURTHER

- Rich source of information besides accidents
- Good archive of data for training

CADS

NEGATIVE SAMPLE PROBLEM

- Model is trying to identify accidents when there is an equal number of non-accidents
- In reality, there are hundreds of thousands more non-accidents than accidents

FUTURE WORK

- Explore Methods that do not require the generation of negative samples
  - Naïve Bayes
  - Clustering in Feature Space
    - Find where accidents tend to cluster
    - Measure how similar or dissimilar a road segment is to the clusters to predict a risk value on that segment

THANK YOU!

Bob Taylor, P.E., PTOE
Chief Technology Officer
Pennsylvania Turnpike Commission
700 South Eisenhower Blvd.
Middletown, PA 17057
Phone: 717-831-7548
Cell: 717-645-1740
robtaylo@paturnpike.com

PTC INFORMATION TECHNOLOGY