Model Mayhem

Benchmarking the Effectiveness of Classification Models on Weather Data

Classification Models

Mult. Regression

Pros: it is a whitebox function, easier to understand

Cons: overfitting the model

Decision Tree

Pros: simple to understand and convey results, easy to visualize

Cons: overfitting very common, poor time complexity

Random Forest

Pros: can deal with unbalanced classes (rare events), and uses dimensionality reduction

Cons: black box method with little control

SVM

Pros: performs similarly to logistic regression when linear separation, handle high dimensional data well

Cons: susceptible to overfitting/training issues depending on kernel

Neural Network

Pros: good for many categories

Cons: requires a lot of data for every category (can't handle rare events)

KNN

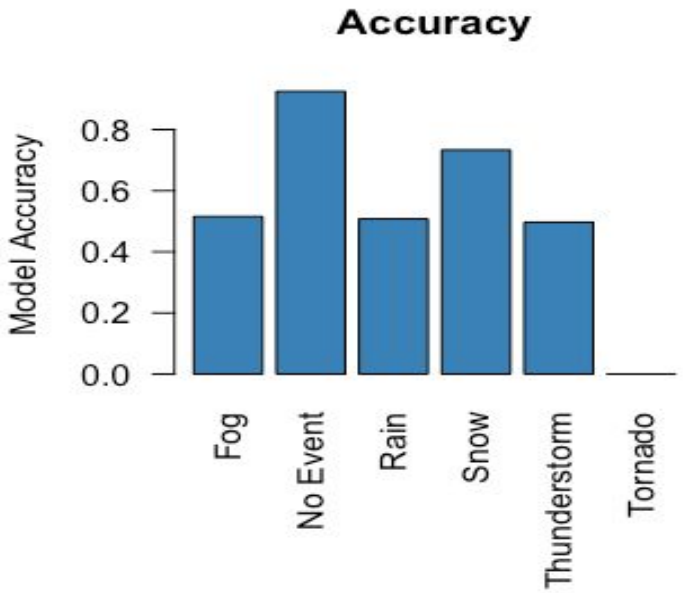
Pros: easy to understand and implement, does not assume probability distribution on the input data

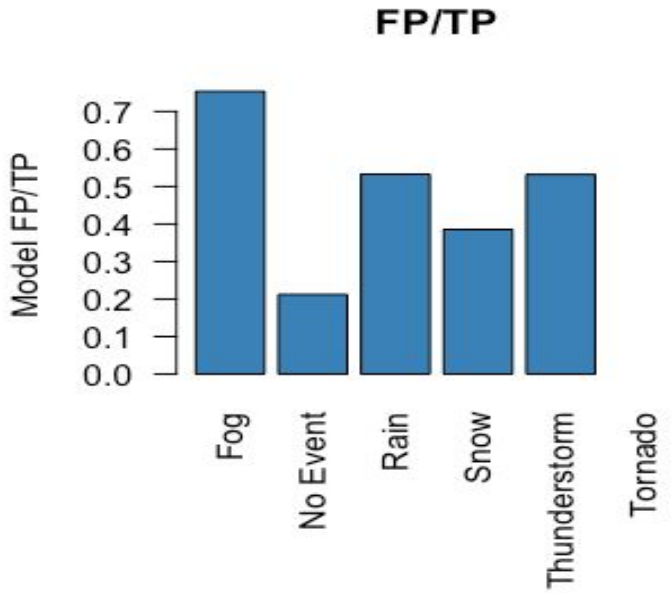
Cons: sensitive in localized data, many dimension could reduce the effectiveness of KNN

Event-Wise Accuracy

	No Event	Fog	Rain	Snow	Thunderstorm	Tornado	Overall
Decision Tree	99.5%	9.4%	0.0%	1.0%	13.4%	0.0%	60.7%
Random Forest	65.0%	0.0%	0.0%	0.3%	51.1%	0.0%	44.1%
Multinomial Regression	92.5%	51.6%	50.8%	73.3%	49.7%	0.0%	76.8%
SVM	NA	NA	NA	NA	NA	NA	NA
Neural Network	0.9%	86.5%	2.0%	25.7%	69.5%	0.0%	13.4%
KNN	87.3%	29.6%	47.7%	67.4%	41.3%	0.0%	71.02%

Best Model: Multinomial Regression





Pred/Ref	Fog	No Event	Rain	Snow	Thunderstorm	Tornado
Fog	609	213	144	65	37	0
No Event	358	17829	2162	312	948	1
Rain	84	694	3265	164	797	1
Snow	86	240	236	1528	26	1
Thunderstorm	38	280	620	13	1789	1
Tornado	6	13	2	1	2	0

Our Learnings

- With a large number of categories most classification models become susceptible to overfitting
- Large sample size is needed for each event, not just overall
- Some models can only work with ordered factors
- Multinomial regression performs the best for our unbalanced classification task