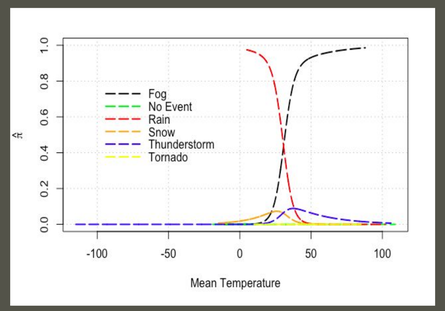
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

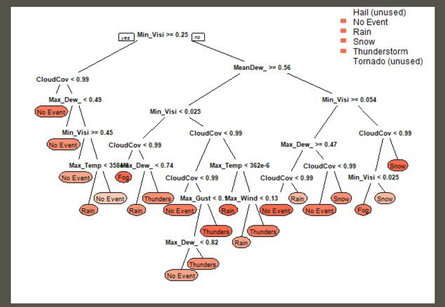
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



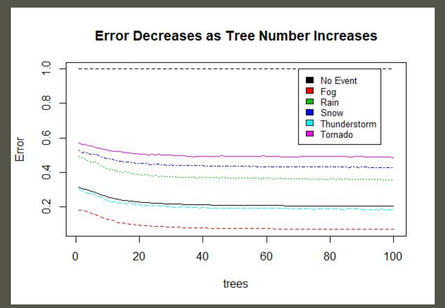
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

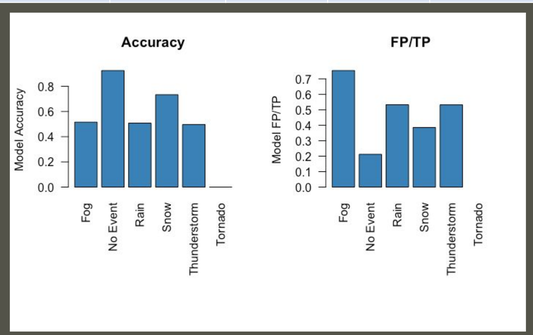
Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

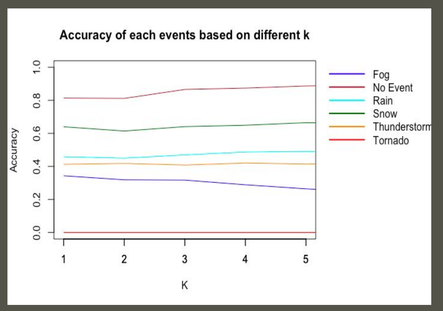
Multinomial Regression

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



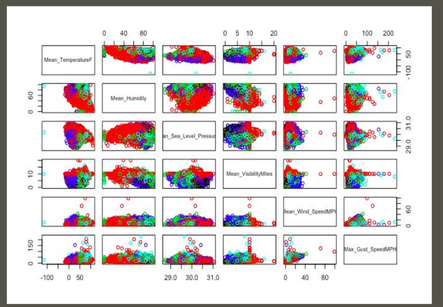
Click headings to further view content

KNN



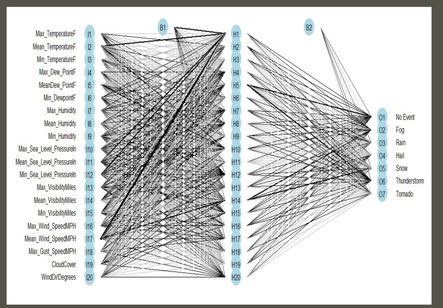
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.

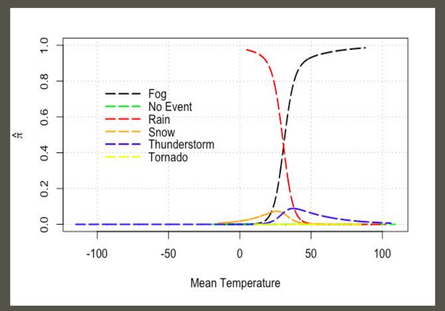
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

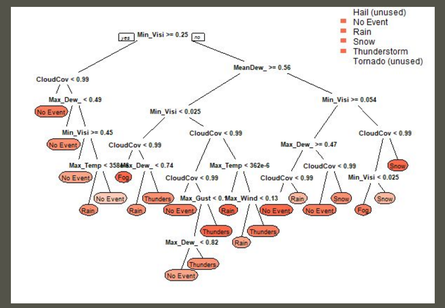
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



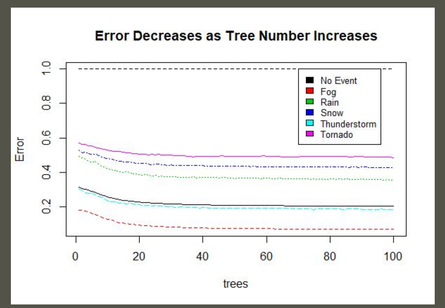
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

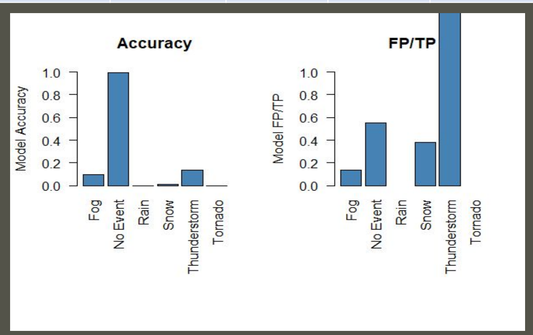
Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

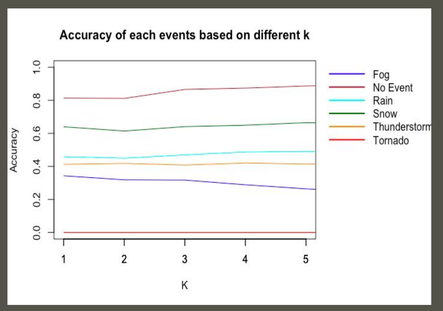
Decision Tree

Pred/Ref	Fog	No Event	Rain	Snow	Thunderstorm	Tornado
Fog	111	225	0	7	838	0
No Event	15	19168	0	1	85	0
Rain	0	5801	0	0	628	0
Snow	0	1447	0	21	615	0
Thunderstorm	0	3117	0	0	482	0
Tornado	0	4	0	0	0	0



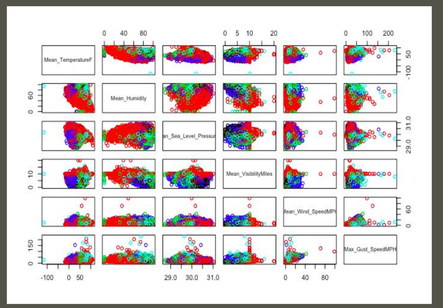
Click headings to further view content

KNN



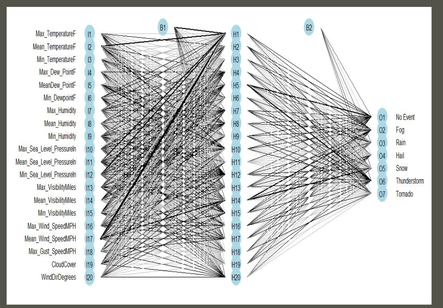
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.

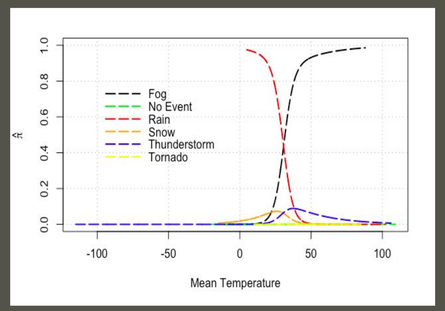
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

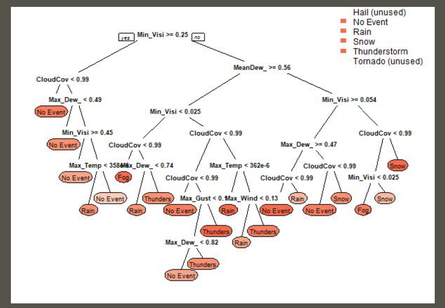
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



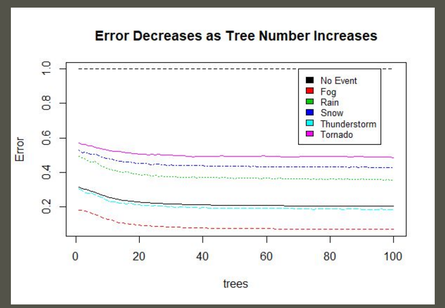
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

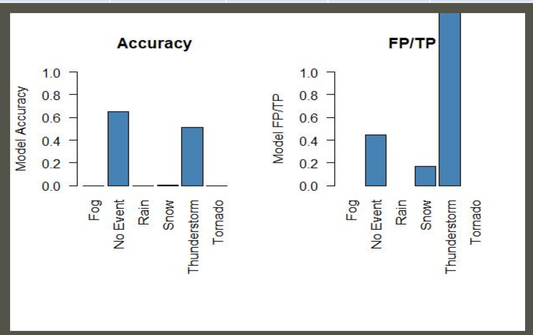
Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

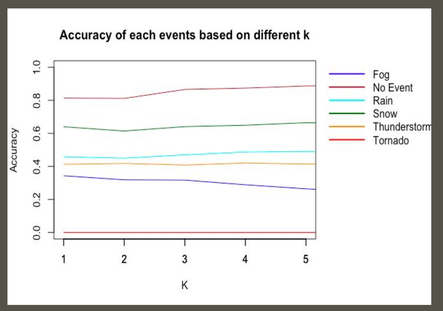
Random Forest

Pred/Ref	Fog	No Event	Rain	Snow	Thunderstorm	Tornado
Fog	0	225	8	0	948	0
No Event	0	19130	7	0	132	0
Rain	0	5799	1	0	629	0
Snow	0	1442	35	4	602	0
Thunderstorm	0	3117	0	0	482	0
Tornado	0	4	0	0	0	0



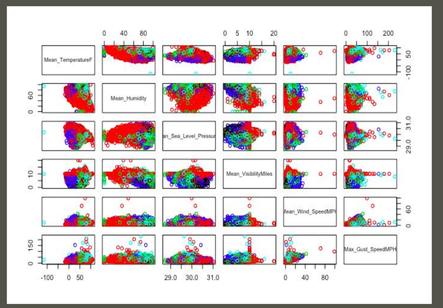
Click headings to further view content

KNN



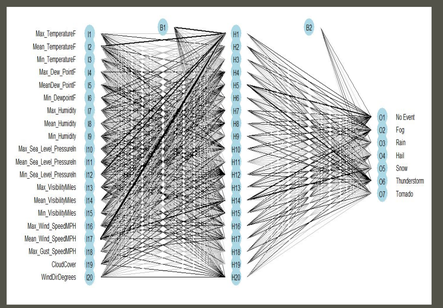
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.

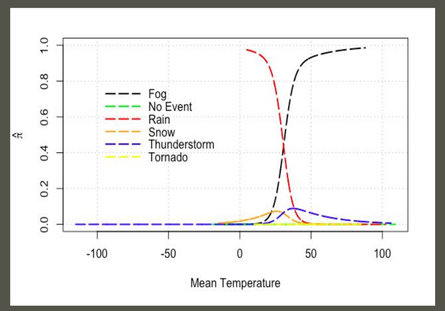
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

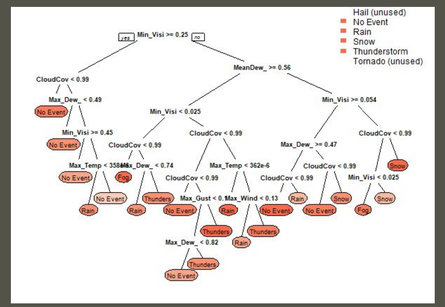
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



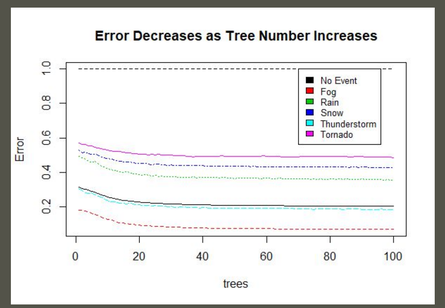
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

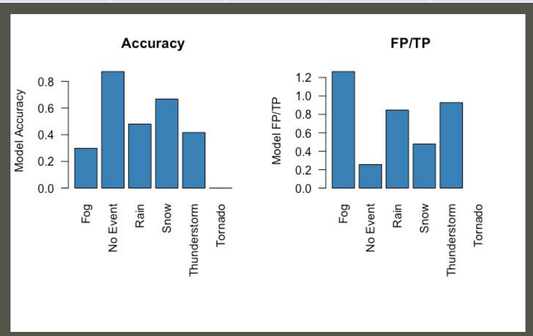
Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

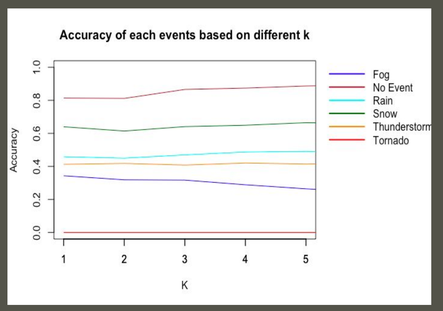
K Nearest Neighbours

Pred/Ref	Fog	No Event	Rain	Snow	Thunderstorm	Tornado
Fog	609	197	147	38	63	0
No Event	435	16831	2267	479	1115	0
Rain	228	1323	3084	157	900	2
Snow	77	384	182	1390	21	0
Thunderstorm	89	534	748	18	1499	2
Tornado	0	0	1	1	1	0



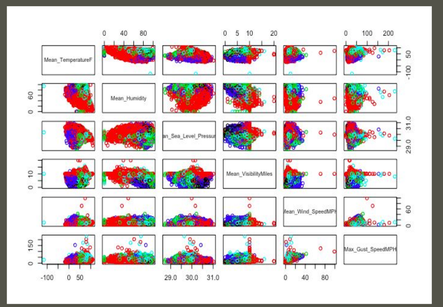
Click headings to further view content

KNN



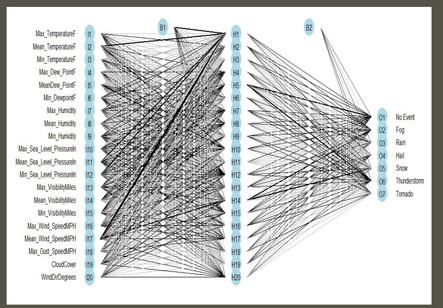
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.

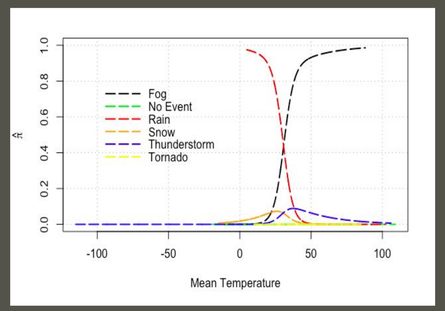
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

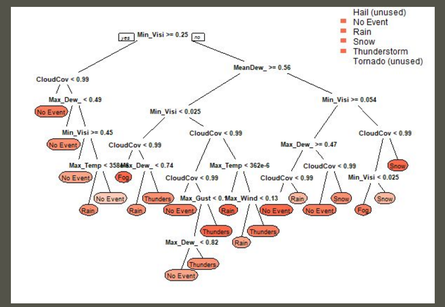
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



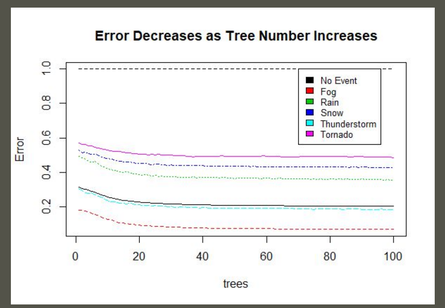
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



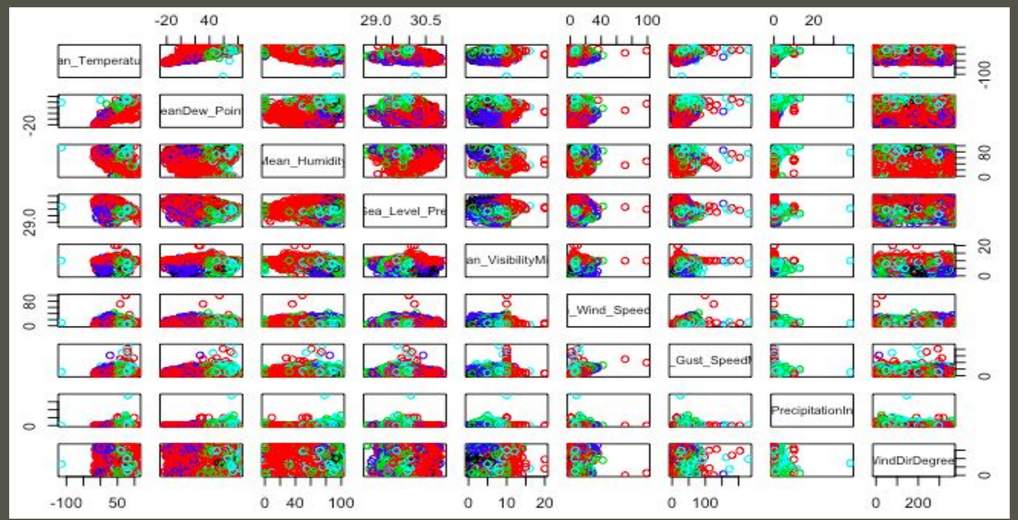
In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

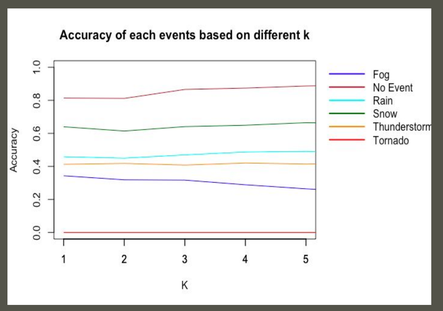
Support Vector Machines



For SVM, we could not run the code in R efficiently due to limited computational resources. Therefore, we did some research to determine our optimum kernel. Firstly, we chose specific features from our dataset as some features might be highly correlated to other features, so we chose the primary features to compare. From the figure above, we suspect that the Radial Basis Function kernel is the optimum as the event groups of data shown above are overlapping with each other. This indicates that the Linear kernel will not work well.

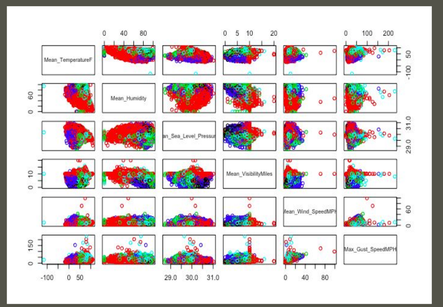
Click headings to further view content

KNN



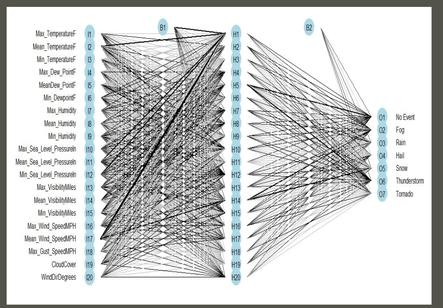
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.

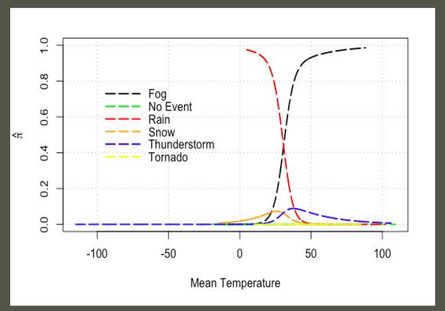
SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Benchmarking the Effectiveness of Classification Models and their Visualizations on Weather Data

Kristen Bystrom
Zhi Yuh Ou Yang

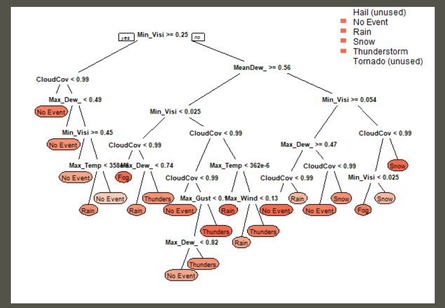
Problem Description: we compare the event-wise accuracy of several classification models at predicting the most extreme weather event that will occur based on the meteorological measurements taken that day. We also evaluate the drawbacks of each model and determine which model performs the best at predicting very extreme events such as tornadoes and thunderstorms, using accuracy and false positive to true positive ratio as our key performance indicators.

Multinomial Regression



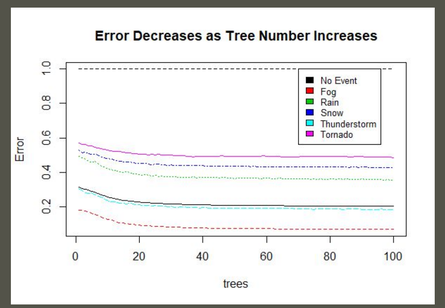
Multinomial logistic regression is used to model nominal outcomes, in which the log odds of the outcomes are modeled as a linear combination of the prectidor variables. We found that multinomial regression has the best overall accuracy and an event-wise accuracy of more than 50% except for the rare event, tornado.

Decision Tree



In statistics, decision trees (also known as classification trees) can be used to predict the outcome of an event. The leafs represent the predicted classes and the branches mark split in the features that determine the predicted class. We found that decision trees tend towards overfitting and can't handle rare cases.

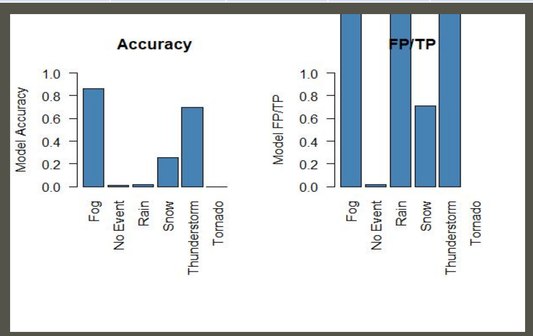
Random Forest



Random forest models are comprised of multiple trees from random subsets of the features. An average of the predictions is taken to produce the final prediction. In the figure to the left, the error reduces until about 40 trees after which it level off. We found that the random forest model was able to deal somewhat ok with the unbalanced categories and was less susceptible to overfitting.

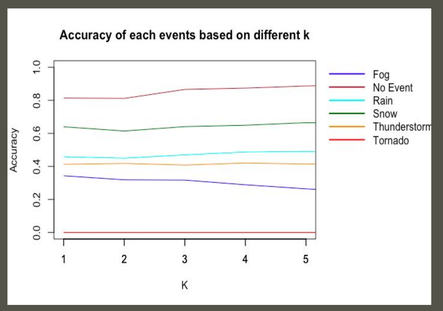
Neural Network

Pred/Ref	Fog	No Event	Rain	Snow	Thunderstorm	Tornado
Fog	1021	0	30	35	95	0
No Event	9609	170	298	343	8849	0
Rain	3724	2	129	0	2574	0
Snow	1203	0	4	535	341	0
Thunderstorm	1051	1	46	1	2500	0
Tornado	2	0	1	0	1	0



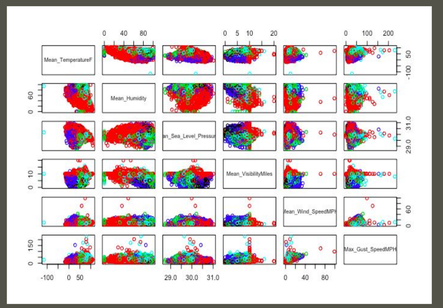
Click headings to further view content

KNN



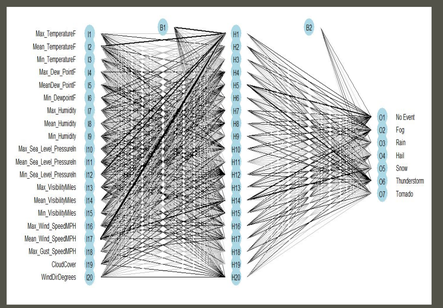
K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure such as a distance function and it is a non-parametric method. We found that choosing k might be a challenging task for this analysis so we plotted a figure based on a range of k from 1 to 20 to indicate the optimum k.

SVM



Support vector machine is a classifier that maximizes the margin between different classes and it is a non-parametric method. We found that our data set represents mostly non-linear relationships between different features, so we suspect it is a radial basis function kernel. However, the computational time was too long to process the tuning of gamma and cost parameters.

Neural Network



Neural networks use one or more hidden layers, combined with some bias and backpropagation, can predict an event. However neural networks require a large amount of data and even though the larger categories have sufficient data, the small categories are too small to be correctly predicted.