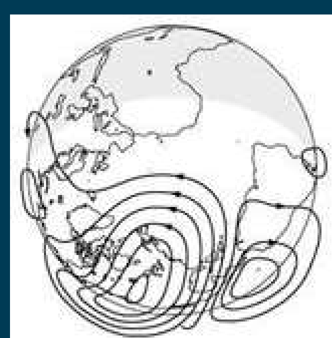
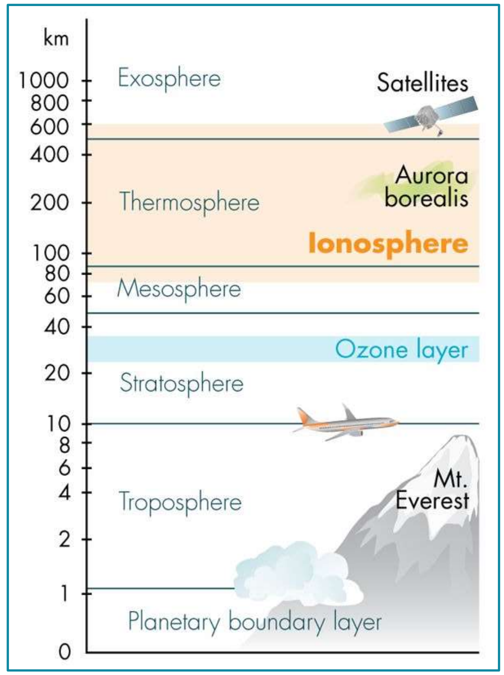
Classification of Radar Returns Using Machine Learning Techniques



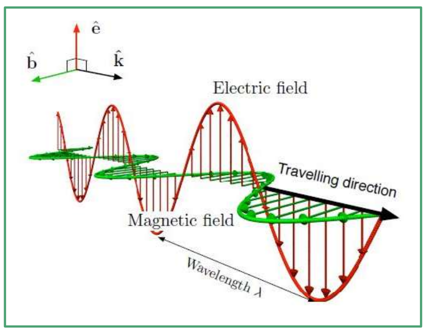
The Ionosphere

- Atmospheric layer with high abundance of free electrons and ions
- Enables radar communication by reflecting radar signals back to Earth
    - The ‘good’ radar signals that are reflected must be separated from the ‘bad’ signals that pass through ionosphere into space



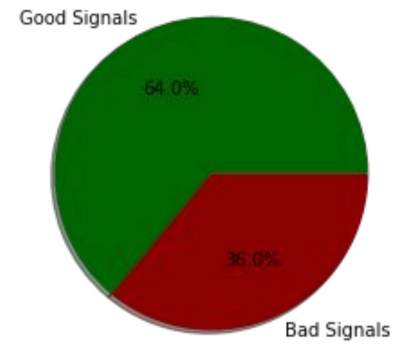
Dataset Used: UCI Ionosphere Dataset

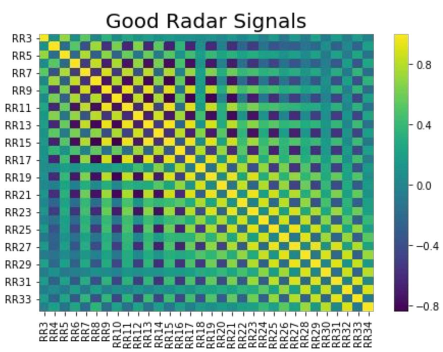
- Consists of radar observations collected in Goose Bay, Labrador
- Received Signals processed by autocorrelation function
- Each pulse number is described by 2 attributes - an imaginary and real part
    - Done to simultaneously describe electronic and magnetic part of electromagnetic wave

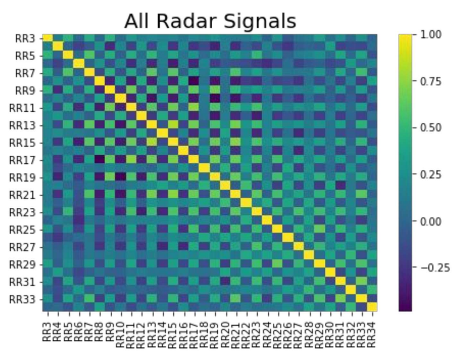


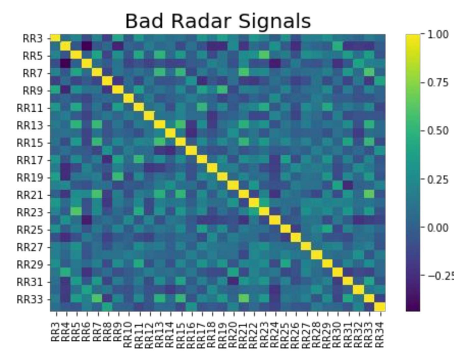
Exploratory Analysis of Data

- 350 total observations consisting of 34 variables
  - All variables scaled between 0-1
- Dropped first 2 columns of data because they did not significantly influence models and contained no useful data









Initial Model Results

Model Type	Training Set Accuracy	5 Fold Cross Validation Accuracy	Runtime (seconds)
KNN Regression	91.14%	24.06%	0.001448
Lasso Regression	36.00%	-7.89%	0.002211
Naive Bayes	84.86%	83.98%	0.002936
KNN Classifier	84.86%	84%	0.005096
Partial Least Squares	82.57%	21.84%	0.007073
Support Vector Machine	94.57%	93.42%	0.007092
Ordinary Least Squares	82.29%	7.93%	0.008913
Decision Tree Classifier	92.29%	89.44%	0.009003
Ridge Regression	82.29%	23.61%	0.009101
Logistic Regression	88.86%	82.02%	0.015642
Random Forest Classifier	98.86%	91.73%	0.019064

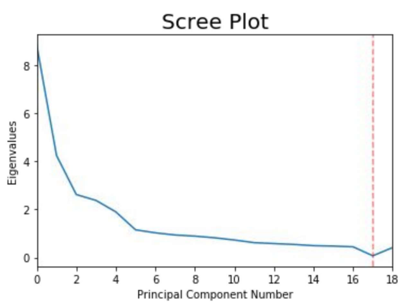
SVM	Actual: False	Actual: True
Predicted: False	113	4
Predicted: True	13	220

Decision Tree	Actual: False	Actual: True
Predicted: False	107	5
Predicted: True	19	219

Random Forest	Actual: False	Actual: True
Predicted: False	125	0
Predicted: True	1	224

Potential Improvements: PCA

- Compressed 34 variables into 17 compressed features using principal component analysis



Model Type	Training Set Accuracy	5 Fold Cross Validation Accuracy	Runtime (seconds)
KNN Classifier	88.29%	84.86%	0.000973
Support Vector Machine	95.14%	93.43%	0.005269
Logistic Regression	86.00%	83.45%	0.007364
Decision Tree Classifier	93.14%	87.99%	0.007526
Random Forest Classifier	99.43%	92.57%	0.016374

- Improvements:
  - Improved overall accuracy of RFC, SVM, DT, and KNN Classifier
  - Reduced runtimes for all models excluding DT which remained relatively unchanged

Potential Improvements: Gradient Boosting

- Most effective parameters found for this model:
  - Estimators: 1000
  - Max Depth: 3
  - Loss Function: Deviance
- Pros:
  - 100% training set predictive accuracy
  - 93.16% 5 fold cross validation accuracy
- Cons:
  - Substantially increased runtime (0.9416 seconds)

Projected Runtime for 1 Year of Observations

- Dataset only representative of 70 seconds of sampling from site
- Site gathers 25 observations every 5 seconds continuously year round
  - 1 second of training set runtime = 125.14 hours for full year of data

Projected Runtimes for 1 Year of Data	
Model	Projected Runtime
KNN Regression	0:14:27
Lasso Regression	0:15:46
KNN Classifier	0:16:07
Ordinary Least Squares	0:23:01
Naive Bayes	0:23:12
Ridge Regression	0:30:03
Partial Least Squares	0:45:02
Support Vector Machine	0:57:02
Decision Tree Classifier	1:07:24
Logistic Regression	1:50:11
Random Forest Classifier	2:24:29

Summary

- The most effective models found were the support vector machine, decision trees and random forest since they produce the fewest type 1 errors of the models tested
- The performance of some of these models can be improved via PCA or gradient boosting
- If runtime is of particular concern the KNN Classifier can be used to substantially reduce runtime but at the cost of predictive accuracy