Detecting Volcanoes on Venus

Unit 3 Capstone

What will be covered

- Topic
- Data and Data Cleaning
- Model Comparison and Selection Process
- Best Model
- Practical Uses
- Weak Points and Shortcomings
- Future Work

Volcanoes on Venus

Volcanism is the most important and well known geologic phenomenon on Venus. Many planetary geologists study venus because of this widespread geological feature. Venus has over 1600 major volcanoes and an estimated 1 million small volcanoes. This number is an estimate as no one has yet to count every small volcanic feature on Venus.

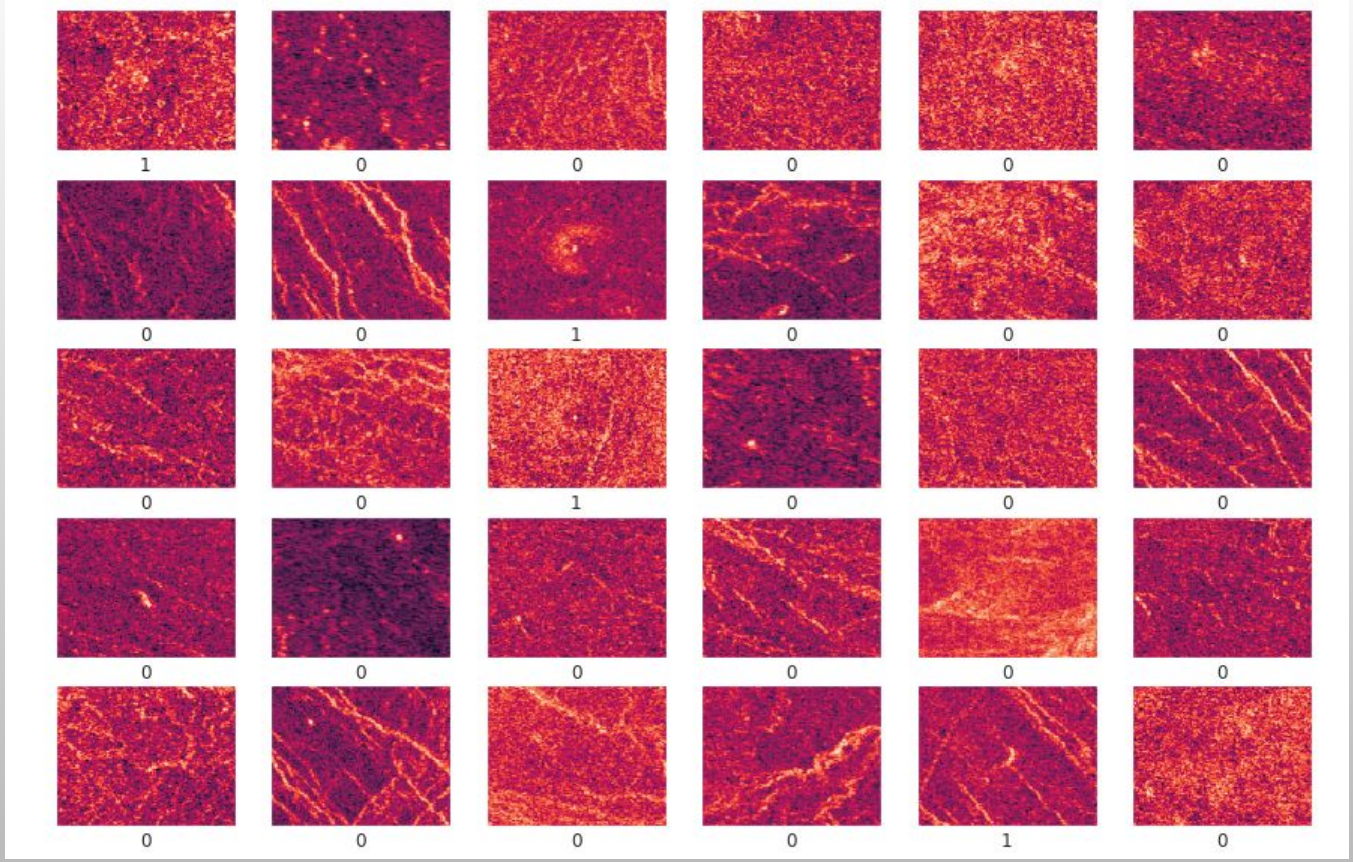
The Magellan spacecraft used Synthetic Aperture Radar (SAR) and was set to map 70% of the Venusian surface. By the end of its mission, Magellan was able to map 98% of the surface.

This capstone project is an attempt to create a supervised machine learning model that can predict volcanoes from images taken from the Magellan mission.

Data

- Kaggle Dataset
- Split train and test (7000 images in train data and 2734 images in test data)
- pixel values
  - Dataset is comprised of pixel values
  - Images - 110x110
  - For the model - each column is a feature and each row is an image
  - 12100 features (all the 110 rows of 110 columns)
- Labels
  - Volcano?
  - Type
  - Radius
  - Number of Volcanoes

Data: Images



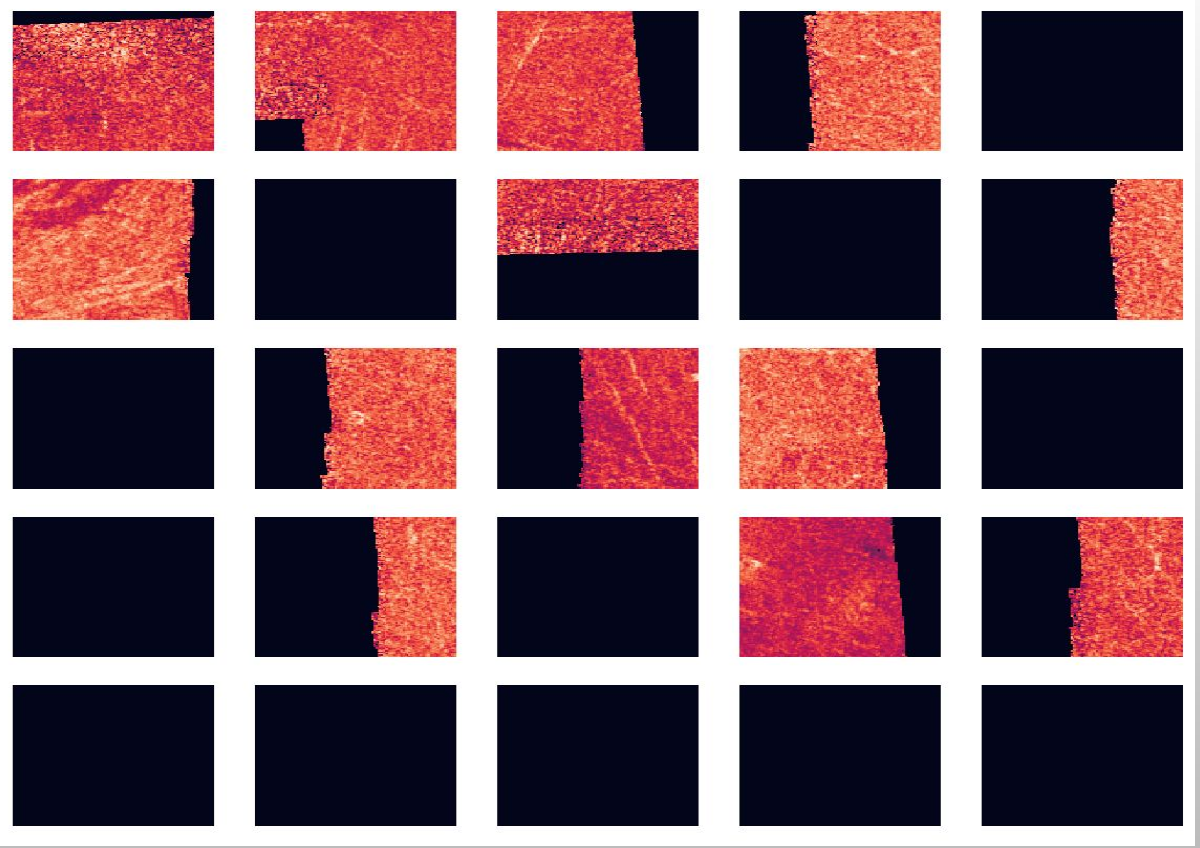
Data cleaning

- For simplicity - binary (instead of number of volcanoes or probability)
  - Detect Volcano - 1 or Do no detect volcano - 0

- Normalized pixel values
  - Normalize pixel values so that each pixel ranges from 0 to 1

- Corrupted Images
  - Remove corrupted images defined by a 0 pixel value

Data: Corrupted Images



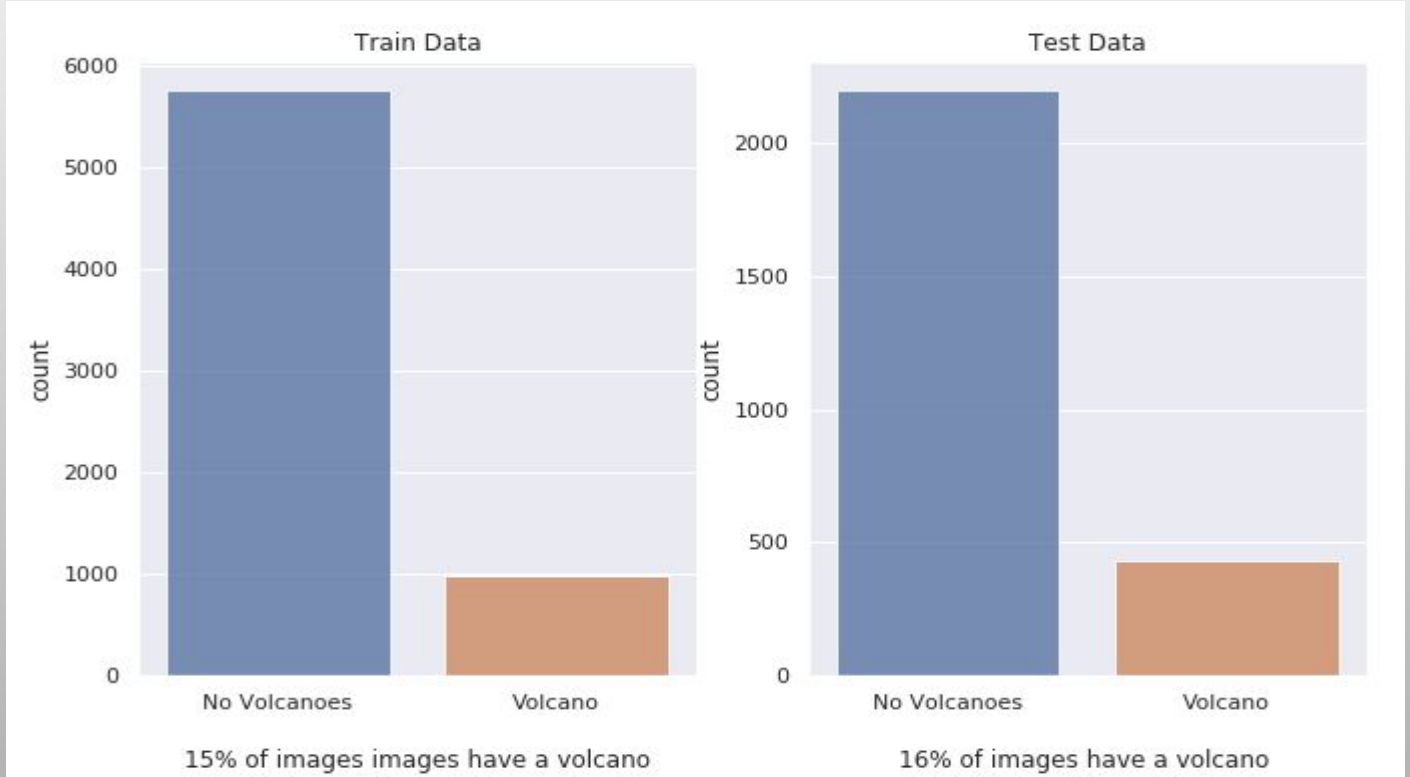
After removing the corrupted images:

Train Data - 6729 Images

Test Data - 2627 Images

Data: Class Imbalance

If my model predicts no volcanoes at all, then my accuracy would still be at 85% on the train data and 84% on the test data.

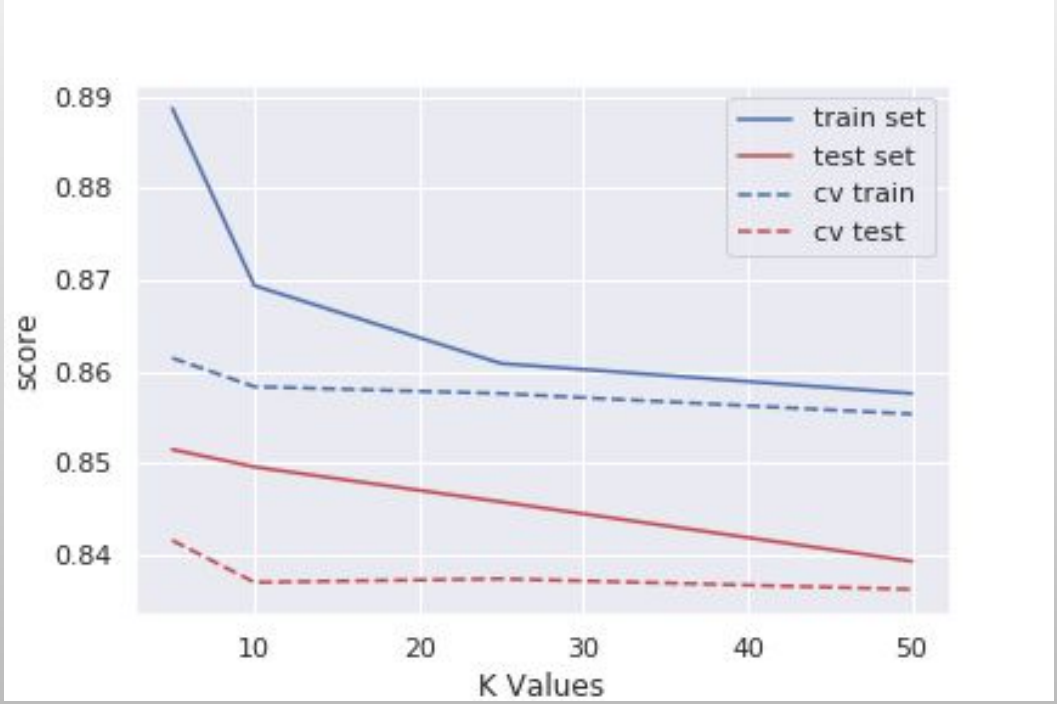


Models

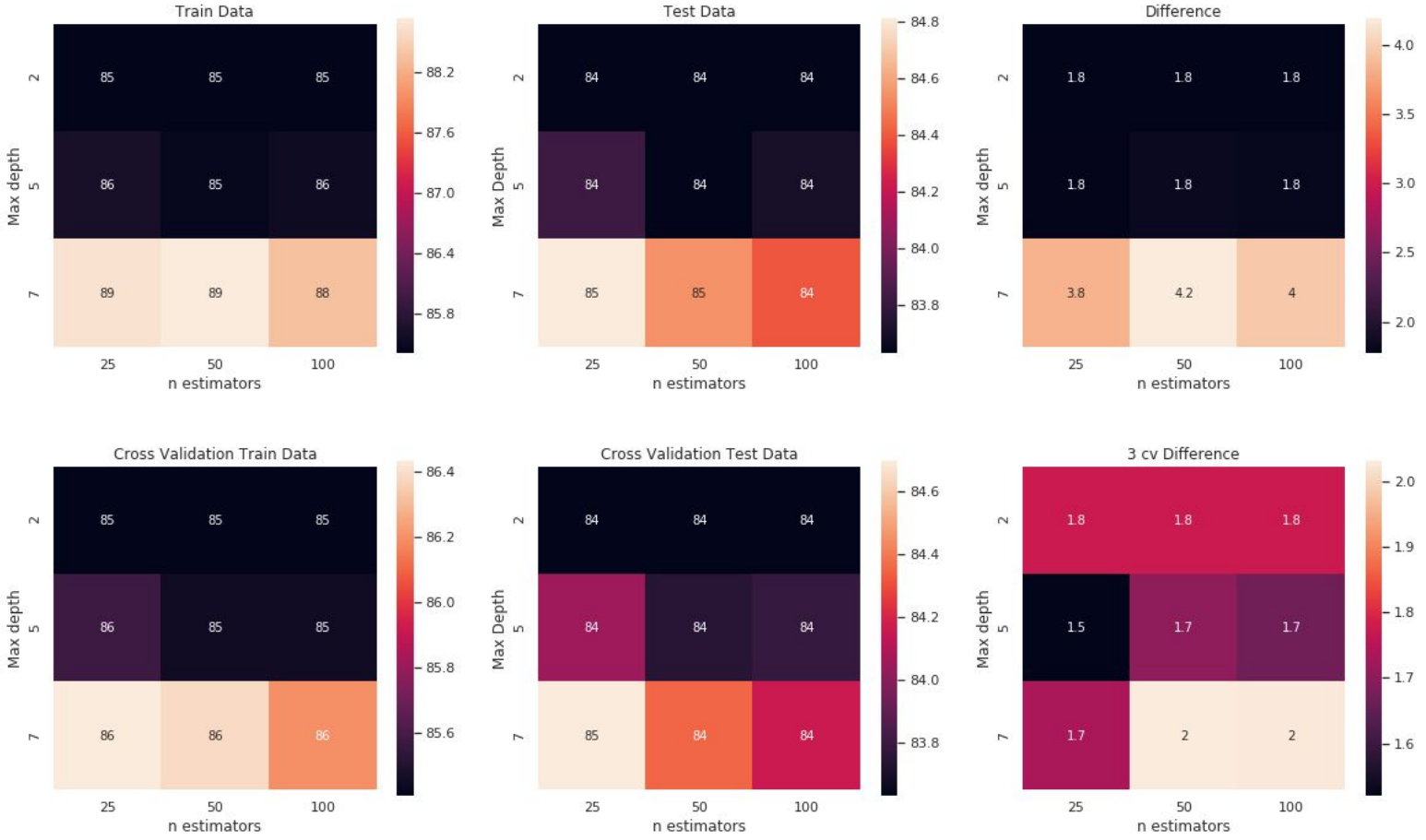
- Classifier - binary outcome
  - K Nearest Neighbor
    - K neighbors
  - Random Forest
    - Max depth
    - Number of iterations
  - Logistic Lasso Regression
    - Regularization parameter - C value
  - Logistic Ridge Regression
    - Regularization parameter - C value
  - Support Vector Machine
    - Kernel - RBF solver (Radial Basis Function)
    - C value
    - Gamma

K Nearest Neighbors

Deciding on the k value:

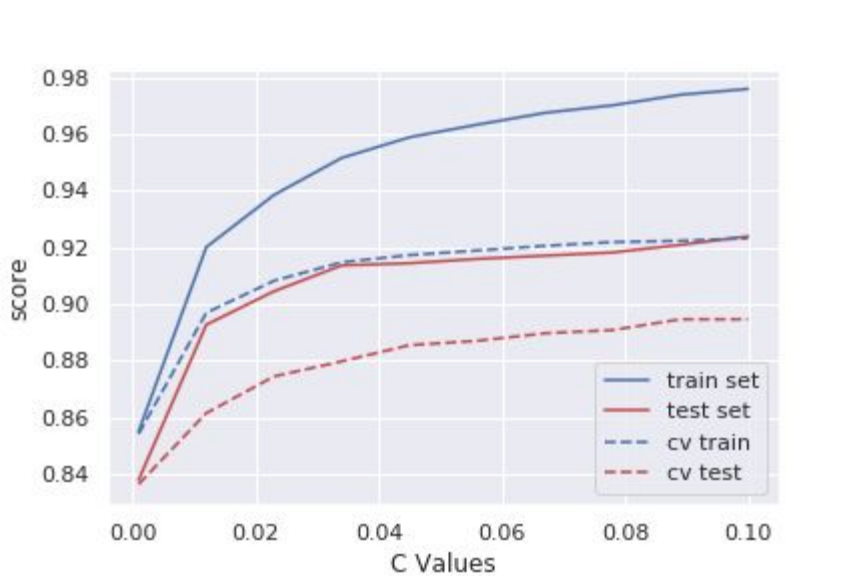


Random Forest

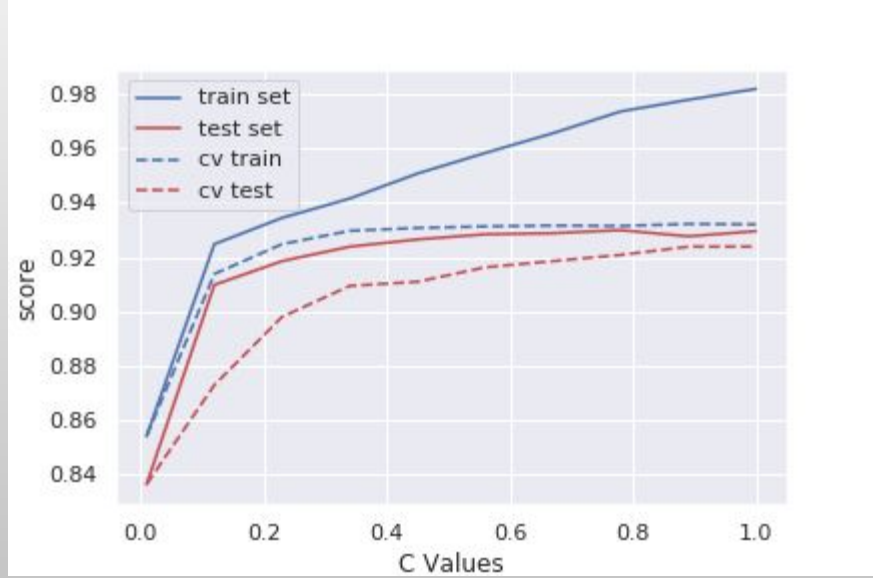


Logistic Regression

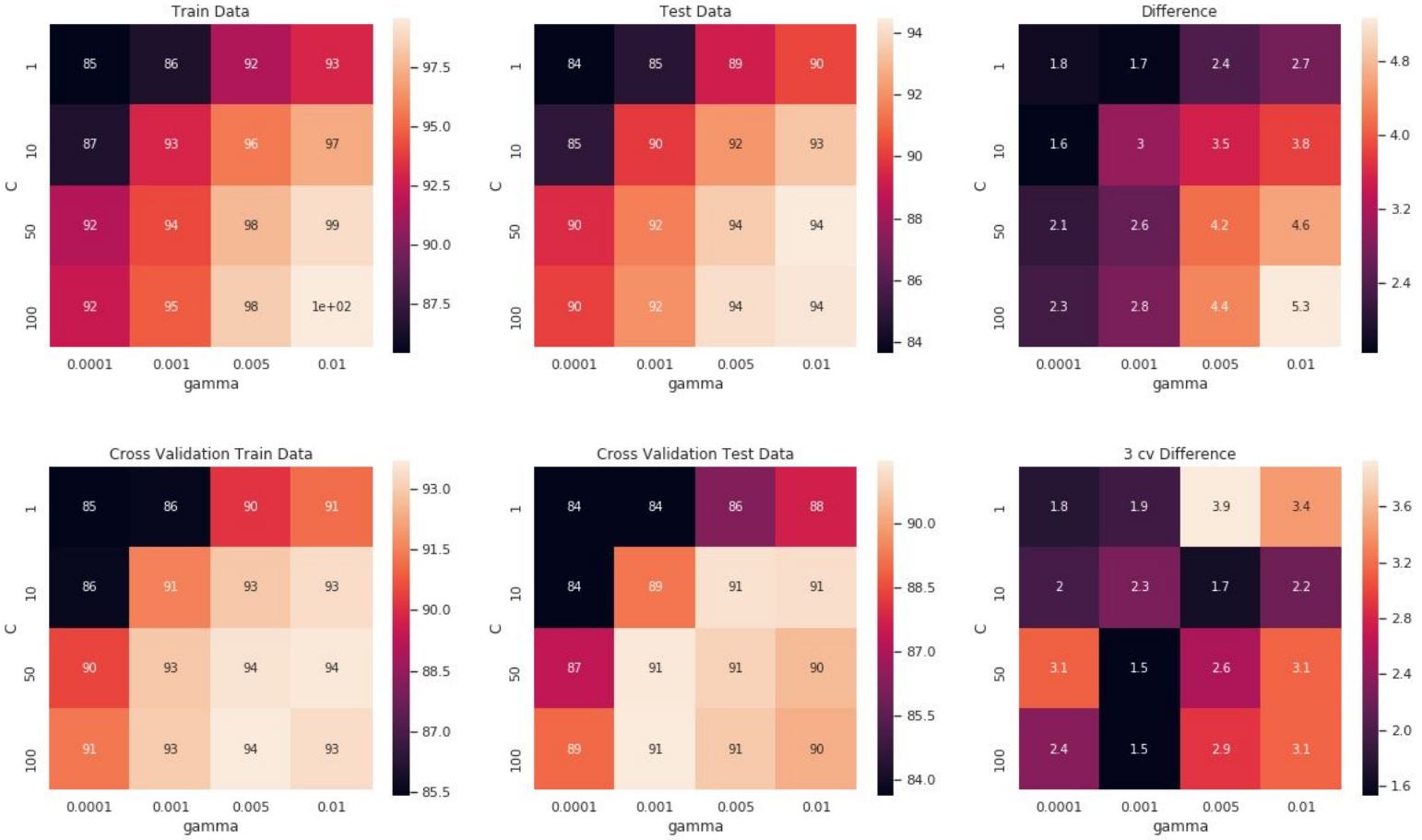
Ridge - L2 Penalty



Lasso - L1 Penalty



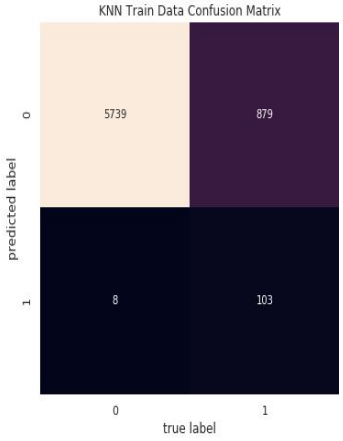
SVM Classifier

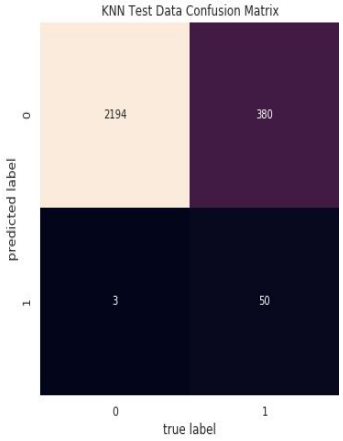


Model Comparison: Accuracy Table

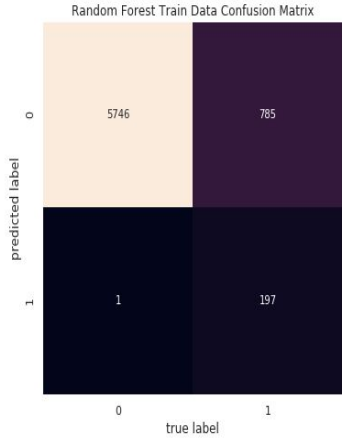
	Test data	Train data	3- fold Cross Validation on Train data	3- fold Cross Validation on Test data
SVM	94%	92%	92.93 +/- 0.41	91.40 +/- 1.01
LR Lasso	94%	92%	92.67% +/- 0.83	90.48% +/- 1.54
LR Ridge	91%	88%	89.18% +/- 0.15	85.84% +/- 0.16
KNN	87%	85%	86.00% +\- 0.21	83.80% +\- 0.03
Random Forest	88%	84%	86.48% +/- 0.72	84.47% +/- 1.03

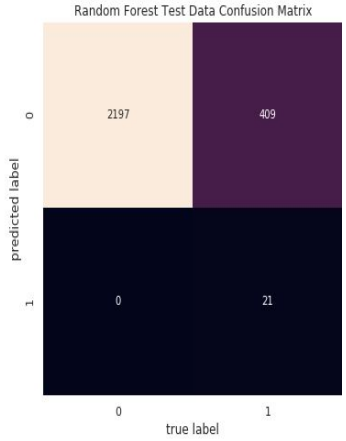
KNN



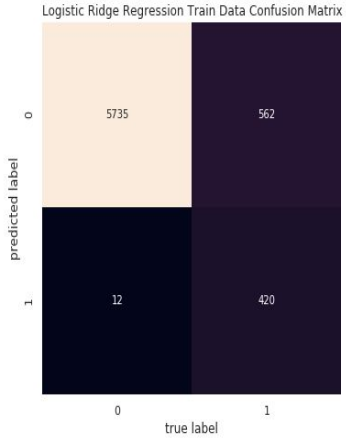


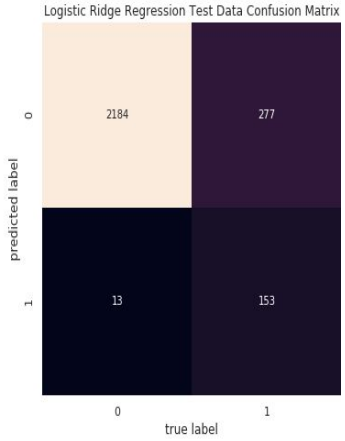
Random Forest



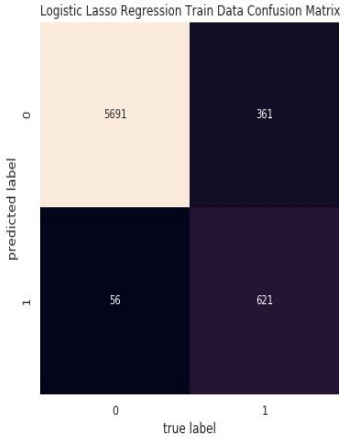


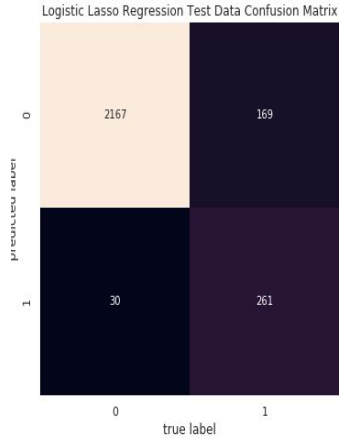
Logistic Ridge Regression



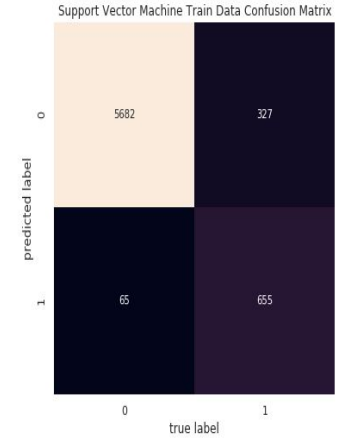


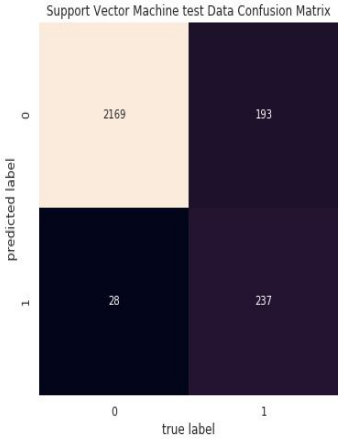
Logistic Lasso Regression





SVM





Model comparison: Type I and Type II errors

Sensitivity: Percentage of positives correctly identified

Specificity: Percent of negatives correctly identified

- KNN
  - Train Data:
    - False Positives: 879
    - False Negatives: 8
    - Sensitivity: 93%
    - Specificity: 87%
  - Test Data:
    - False Positives: 380
    - False Negatives: 3
    - Sensitivity: 94%
    - Specificity: 85%
- Logistic Lasso Regression
  - Train Data:
    - False Positives: 361
    - False Negatives: 56
    - Sensitivity: 92%
    - Specificity: 94%
  - Test Data:
    - False Positives: 169
    - False Negatives: 30
    - Sensitivity: 87%
    - Specificity: 93%
- Support Vector Machine
  - Train Data:
    - False Positives: 327
    - False Negatives: 65
    - Sensitivity: 91%
    - Specificity: 95%
  - Test Data:
    - False Positives: 193
    - False Negatives: 28
    - Sensitivity: 89%
    - Specificity: 92%

Practical Uses

Mapping the location, size, and other information about each volcano would allow for more advanced geophysical analysis.

Information on the general distribution and clustering of volcanic features on Venus is important in understanding the geologic evolution of the planet and allows scientists to better understand features such as the local tectonic structure, heat flow, or volcanic eruption mechanics.

Weak points and shortcomings

- Class imbalance
- Binary Outcome is limiting
  - no probability, never 100% positive
  - Does not allow for real mapping or location
- Overfitting
- Model takes a long time to run
  - Model is limited since even this simple model takes between 30 minutes to an hour to run
  - More complicated models would take much longer to run
- Lost information in corrupted images

Future Work

- More than a Binary outcome
  - Predict number of volcanoes in image
  - Size of Volcano
  - Probability
- Work with Corrupted data
  - Do not want to lose valuable information
- Mapping Volcanoes on Venus
  - Mapping
  - Clustering