Detecting Volcanoes on Venus via Classification (Where are the Volcanoes?!!)

STAT 432 Final Project
Team Mamamia!
Di Ye (diye2), Hao Wang (haow2),Hannah Pu (chpu2)
November 17, 2018

Final Report

INTRODUCTION

Data Source Information

The data was downloaded from Kaggle, which is originally from NASA's Magellan spacecraft database. (https://www.kaggle.com/amantheroot/finding-volcanoes-on-venus/data)

Data Description

9734 images were captured by the spacecraft and converted to pixels (110 x 110, from 0 to 255), where every image is one row of 12100 columns (all the 110 rows of 110 columns). Images can contain more than one volcanoes or maybe none. The 9000+ images are separated to four datasets (file names : *train_images*, *train_labels*, *test_images*, and *test_labels*):

Image Dataset (*train_images* and *test_images*)

Train_images : 7000 images as train data with 12100 variables;
Test_images : 2734 images as test data with 12100 variables; All the variables (V1 to V12100) correspond to the pixel image, 110 pixels * 110 pixels = 12100 pixels.

Label Dataset (*train_labels* and *test_labels*)

A summary of the variables in both *train_labels* and *test_label* datasets is listed down below:
1. *Volcano?* : If in the image there exists at least one volcano, the label is 1 (Yes). Otherwise, the label is 0 (No). (If Volcano? equals 0, the following three categories would be "NaN"). 2. *Type* : 1 = definitely a volcano,2 = probably, 3 = possibly, 4 = only a pit is visible
3. *Radius* : Is the radius of the volcano in the center of the image, in pixels?

4. *Number Volcanoes* : The number of volcanoes in the image.

Literature Review

In the Kaggle, the data analysis of the project is done in Python. People have already had vivid data visualization and exploratory data. Different methods have been used, such as Convolutional Neural Network (CNN) and VGG Neural Network for deep learning. People have reached the 95% accuracy.

Scientific Goal

For this project, we will focus mainly on predicting whether each image has a volcano or not. In addition, if the classification prediction goes well, we will also construct models to predict the number of volcanoes in the images. We aim in constructing different classification models and choosing the best model to predict whether there exists a valcano in each image. Identifying valcano through IT technology would increase the efficency of space exploration and safty of the crews.

EXPLORATORY DATA

The first 6 observations of *train_labels*

```
##   Volcano. Type Radius Number.Volcanoes
## 1        1    3  17.46                1
## 2        0  NaN    NaN              NaN
## 3        0  NaN    NaN              NaN
## 4        0  NaN    NaN              NaN
## 5        0  NaN    NaN              NaN
## 6        0  NaN    NaN              NaN
```

The first 6 observations of *test_labels*

```
##   Volcano. Type Radius Number.Volcanoes
## 1        0  NaN    NaN              NaN
## 2        0  NaN    NaN              NaN
## 3        1    1  17.00                1
## 4        0  NaN    NaN              NaN
## 5        1    3  15.13                1
## 6        0  NaN    NaN              NaN
```

After exploring the datasets, we found only labels have NaNs. We have set the those values to 0, an insignicant value in our dataset.
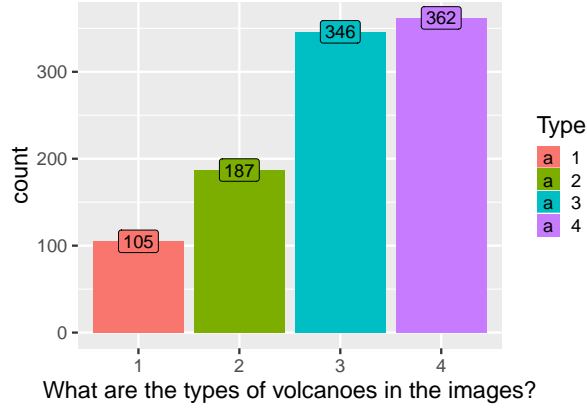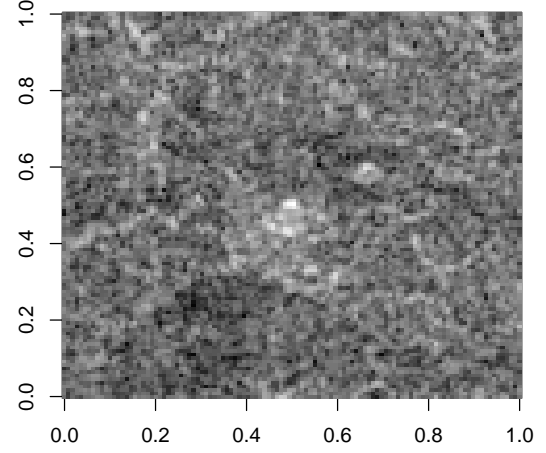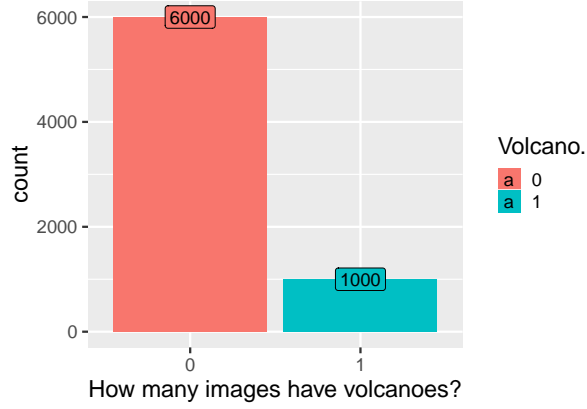

Figure 1: Obs = 1 Type: 3 Radius: 17.46 Number of Volcanoes: 1

METHODOLOGY

Models we used:
Lasso Regression
ElasticNet
Neural Network
Neural Network was used in the project for seeking better classfication result. The raw data train_images and test_images was converted into numeric variables and reshaped into 3D array with dimension $7000 \times 110 \times 110$ and $2734 \times 110 \times 110$ respectively. The pixels were shrinked from 0-255 to 0-1 by dividing each pixel by 255. The target label (Volcano?) was converted into categorical variable for classification. The neural network used several convolutional dense layers for classification. The neural network model yielded a satisfying classification result.
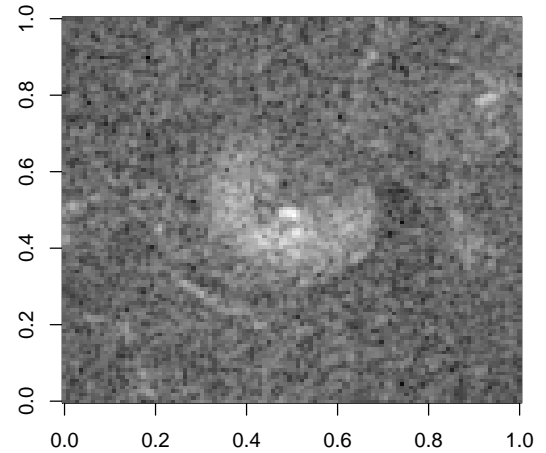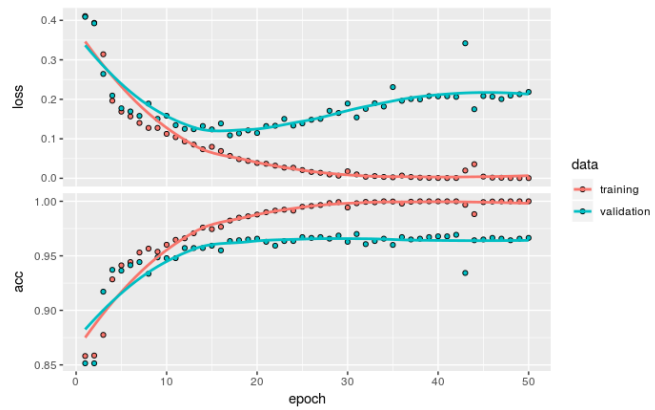

Figure 2: Obs = 10 Type: 1 Radius: 22.02 Number Volcanoes: 1

	No	Yes
No	2270	88
Yes	30	346