Regression Using Support Vector Machines and caret

Your results should be delivered as an HTML webpage generated using R Markdown. Make sure to include all code and results along with the answers to the questions. Provide text to describe your methods and results. This should read like the Methods and Results sections of a paper.

Grading Criteria

- Correctness and completeness of code (16 Points)
- Description of process and results (12 Points)
- Answer to questions (4 Points)
- Webpage formatting (4 Points)
- Map output (4 Points)

In this assignment, you will predict forest percent canopy cover within 30 meter Landsat pixels using four predictor variables: normalized difference vegetation index (NDVI), brightness, greenness, and wetness. The four predictor variables were derived from the Landsat bands. Training data (**training.csv**) and validation (**validation.csv**) have been provided as CSV files. You will also predict out to the image pixels. A small extent of a raster grid stack has been provided (**pred.img**) that contains all four predictor variables. You will make predictions with the support vector machines (SVM) algorithm using the **caret** package. Since you only want to make predictions within the extent of forests, a forest mask has been provided (**forest_mask.img**) to mask the result.

Complete the following tasks:

- Create four scatter plots to visualize the relationship between percent canopy cover and the four predictor variables using **ggplot2** and the training samples.
- Use the training data and **caret** to generate a model using SVM (method = "svmRadial"), a tuneLength of 10, and optimize relative to the RMSE metric. Use 5-fold cross validation to tune the model.
- Predict to the validation data and calculated MSE and RMSE from the result.
- Predict to the raster stack. Note that you will need to rename the bands ("ndvi", "Brightness", "Greeness", "Wetness").
- Mask the result relative to the forest raster mask.
- Create a map of the result using **tmap**. Use an appropriate palette and provide a legend and title.

Q1: Provide a discussion of each of the four scatter plots. Do you see a relationship between percent canopy cover and each specific predictor variable? Describe the relationship. Which variables appear to have the strongest relationship with percent canopy cover?

Q2: What is the reported MSE? What are the units of MSE for this prediction?

Q3: What is the reported RMSE? What are the units of RMSE for this prediction?

Q4: Provide an interpretation of RMSE? Is this a strong prediction or is there a lot of error or uncertainty?