Assignment 1: Logistic Regression & Linear Regression with Python

This assignment is worth 100 points total. You must show code to back up your results e.g. if you report just a number, you will get zero credit. Submit your code in a Jupyter Notebook. Include your name in the file name (e.g. "Assignment 1 – [first name] [last name]"). Clearly label each section of your code that responds to a particular question. There should be a labeled section for Question 1, 2, ...etc. For example, you could call the first section "Question 1: Reading in datasets".

Due Date: July 3rd at 11:59PM (note this is a later date than on the syllabus)

<u>IMPORTANT:</u> Show all code for each step in every problem.

Part 1

In Lab 1 and 2, you worked with housing data. In Part 1 of this assignment, you will revisit the same dataset. You may copy some of your lab code over (e.g. reading / merging datasets). Instead of using standard OLS, in this assignment, you will be using Lasso and Ridge regression.

1. Following the same logic as Lab 1, read in train_2017.csv and properties_2017.csv. Join the two datasets together based off parcelid and use an inner join. You should be able to reuse your code for this step from Lab 2.

2. Split the dataset into training / validation using the standard 70 / 30% split.

3. Do any data cleaning you need to before proceeding to the remaining steps. It is your job to figure out what cleaning you may need to do. (15 points)

4.Use as many raw inputs as you think necessary and train a lasso regression model using the training dataset. Explain your initial selection of

variables. This problem is meant to be vague – you decide what set of variables you want to start with – you just need to provide a reasonable explanation as to why you might want to exclude certain variables vs. keep others. Imagine you are explaining this to your data science manager at work.

a. Provide initial set of variables and corresponding explanation as described above (5 points)

b. Tuning for the optimal parameters is often needed in training machine learning models. One way to do this is to perform k-fold cross validation on the dataset. K-fold cross validation is the process of dividing a dataset into k subsets. A model is then trained using k – 1 combined subsets and validated on the left-out dataset. This is repeated k- times so that each subset is left-out for validation exactly one time. Use k = 5 and perform cross validation on the training set to tune for the optimal value for the regularization parameter. Report this value and show the code you write to obtain this value. (10 points)

5.Train a ridge regression model using your initial set of variables (also using 5-fold cross-validation). Compare the performance against the lasso model. Is one better than the other? Explain. (5 points)

6.Using any or all of the raw variables, try creating new features and testing these in the model. (15 points total)

a. Define how you would measure "performance" from a technical perspective

b. How would you explain a) in a non-statistical / non-technical way?

c. What's the best model performance you can achieve? Try at least 4 – 5 features.

Part 2

In Part 2 of this assignment, you will be predicting customer churn using logistic regression. Customer churn is defined as when a customer decides to stop being a recurring customer. Canceling your Netflix account is an example of customer churn. Predicting customer churn is a common task for businesses because if they know what customers are more likely to quit their services / products, they can make changes or target those customers for retention. This can ultimately drive higher revenue and potentially happier customers.

1. Read in the churn_data.csv dataset. Report the following (6 points):

a. Number of rows and columns

b. List the columns with missing values

c. Get a frequency count of the "target" field

2.Split the dataset into training / validation. Use a 70 / 30 split (70% for training, 30% for validation). (2 points)

3.Do any necessary cleaning / processing that needs to be done before modeling. It's your job to investigate and figure out what needs to be done here. (7 points)

4.Feature selection can be an important part of the modeling process. In a business context, it's useful because it can make a model less complex, which potentially means less maintenance and less room for error. This is useful, for example, if the sources feeding into the data are coming from different pipelines. In logistic regression, one way of performing feature selection is to use the L1 penalty, similar to Lasso regression. Using this method (20 points total):

a. Build a logistic regression model using all the raw inputs in the dataset. Provide a list of the variables that have coefficients shrunk to zero (make sure to show your code)

b. Next, try varying values of C. Like in Part 1, use k-fold cross validation

on the training set to optimize this parameter value. Report this parameter value. Using this optimized model, calculate the precision, recall, accuracy, and AUC for the model based off the training set and validation set.

c. Re-do part b), but with no penalty. What do you notice?

d. Re-do b), but this time use L2 penalty.

5. Try building additional features and test those in the model. What's the highest performance in the model you can get by engineering different features? You can also perform feature selection here. The goal of this problem is to achieve the best "performance" possible. Show all of the code and different sets of features you try before arriving at a final set. Additionally, define and explain your definition of "performance". Try to test at least 3 - 4 new features. (10 points)

6. What are the advantages and disadvantages of using logistic regression on this dataset? (5 points)