Megan Cusey

November 25, 2019

MDS 556

Mid-Term Short Essay

**Someone tells you that they applied a linear regression model to a dataset with 100 features and 1,000 rows of data. Their goal is to determine whether or not a new discount offer is likely to boost customer interest in their product. They feel pretty good about the fact that their regression analysis predicted a rise in the number of customers backed by a value for R^2 of 0.54. You are not so sure.**

**Write two or three paragraphs in which you identify the issues that might be in play, specific options you might have for handling this situation in a better way, and how you might go about choosing a course of action that would deliver a better result for this client. Post your essay here.**

The first issue I notice in the situation above is the number of features compared to the number of rows. If you use the rule of 20 observations per feature, the model would be violating this rule of thumb by double. R-squared might not be too reliable since R-squared only increases as the number of features increase. R-squared adjusted would be a better measure of how well the features capture the potential interest in a product when a discount is applied.

In addition, I believe it’s unlikely that each of those features have a linear relationship with the target variable, are truly independent of each other, and there aren’t some instances of multicollinearity. While linear regression has been known to do okay with some of these assumptions violated, the modeler might be able to find a better technique that doesn’t have the assumptions as the basis for the machine learning algorithm.

If this was my project, I would return to feature engineering and figure out how I can limit the number of features in my dataset. Some of the features might be able to be combined, others may be determined to be not important. Business domain knowledge may be inserted in the form of manual variables in order to increase predictive power. In this case, not as many features are required to have a high-quality model. Feature selection techniques such as stepwise forward and backward selection can assist in limiting the features to those that have a significant impact on the target variable.

Once I was content with the features selected, I would split my data into test and train sets. Then apply a series of modeling techniques to the data. Review the results and see if there’s any information in the model outputs that may be applied to the features in order to improve model quality. If there is some information that can be learned and applied to the features, then an iterative approach to model building could be a good approach. That is, return to feature engineering or feature selection, apply leverage the knowledge recently learned, and then revisit model building.

Once the best model(s) have been built with special attention to limiting model complexity, I would evaluate the model further. How much do the training results vary on the test results? Which model reduces RMSE? Which model produces a higher R-Squared adjusted? Do the residuals have a mean of about zero? Is there any evidence that the model might be systematically over or under predicting data points? As you can see, having a somewhat satisfactory R-Squared is not the only tool in evaluating a good model. The above questions should also be considered.