Regression Models

Mary Manz\*

MATH637 - Math Techniques in Data Science

University of Delaware

Newark, DE 19716

May 21, 2021

1. Summary

It is not always obvious which machine learning models will best suit one’s data. And so, I predicted NFL game scores using the starting players’ PFF grades from both teams using several different models. I would then select whichever model had the best performance metrics to continue my analysis. Since my target variables are continuous, I used regression models. In this whitepaper, I discuss the following machine learning models: decision tree regressor, random forest regressor, linear regression, ridge regression, lasso regression, support vector regression, and gradient boosting regressor. I also discuss the performance metrics used to evaluate each model.

1. Decision tree regressor

A decision tree regressor begins with a root node, which represents the entire data set. This node is then split into multiple sub-nodes, each of which are split into multiple sub-nodes, etc. If a sub-node does not split into further sub-nodes, it is considered a leaf. If it does split into further sub-nodes, it is considered a decision node. From the root node and sub-nodes, the data is split based on the different features used in the model. For features that are numeric, the feature data will be considered in two categories: above a threshold and below a threshold. The threshold is determined by the value that results in two groups with the highest pre-determined metric, say standard deviation. The feature to use in the first split is determined by which feature, when used to split the data, results in groups of data (output) with the highest standard deviation. This process continues until you are only left with leaves.

1. Random forest regressor

A random forest regressor model uses the average of several decision tree regressor models. The number of decision tree regressor models can be defined. And each decision tree regressor model uses slightly different training data from “bootstrapping.” That is, the training data is randomly selected, but each sample selected is from the original pool of training data. The resulting sample size is the same as the training data, but it will likely contain redundant samples.

1. Linear regression

A linear regression model is a simple equation consisting of each feature with an assigned weight that it is multiplied against, all of which is summed, along with a fixed (bias term) value. The result is the prediction. The model’s goal is to select weights for each feature (or coefficient) that will reduce the cost function the most. Typically, the cost function is defined as the MSE (mean squared error), which is the average of the squares of errors, which are the difference between the predicted and actual values. The cost function is reduced using the method of gradient descent, which “consists of looking at the error that our weight currently gives us, using the derivative of the cost function to find the gradient, and then changing our weight to move in the direction opposite of the gradient” [5].

1. Ridge regression

Ridge regression is a good model to use when one does not have a large sample size or has more features than samples. It is also like (multiple) linear regression. It effectively has the same equation as previously discussed for linear regression except there is another bias variable, or the “ridge regression penalty” [6]. It is computed by “multiplying lambdaby the squared weight of each individual feature,” for which weight is referred to as ‘coefficient’ in statistics [6]. The value for lambda is whichever value gives the lowest variance, determined using cross validation. Ridge regression is favorable when multicollinearity exists amongst the features. It essentially helps reduce features that are highly correlated with one another from being treated as completing independent features. Imagining that highly correlated features were instead the same exact features, then a linear regression model, which tries to limit the residual values for each feature used, would give more importance to the duplicate features. As such, ridge regression tries to reduce the importance given to highly correlated variables.

1. Lasso regression

Lasso regression is the same as ridge regression except the penalty is computed by multiplying lambda by the absolute weight of each feature. This difference in penalty calculation allows the model to reduce the weight of some features to zero.

1. Support vector regression

Support vector regression is like linear regression except the goal is to find the line that instead of minimizing the residuals, maximizes the margin. This line is referred to as the hyperplane. The margin is the distance between the hyperplane and the closest data/observation points, or support vectors, on either side of it.

1. Gradient boosting regressor

A gradient boosting regressor model is created by first using a “weak learner,” which “are models that perform slightly better than random guessing” [12]. Typically, a form of a decision tree model can be used as the weak learner. Then the results of the weak learner are evaluated. A lower weight is assigned to samples with the best predictions and a greater weight is assigned to samples with the worst predictions. A new model is then fit to this weighted data set and the process repeats. Each subsequent model is better at predicting samples that the previous model had difficulty with. The final gradient boosting regressor model is the “the weighted sum of the predictions made by the previous tree models” [12].

1. Model metrics

Since model performance metrics are not as straightforward for regression models as they are for classification models in determining the best fit model, I decided to evaluate my models across several performance metrics. These metrics were: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared.

MAE is “the average absolute difference between observed and predicted outcomes” [13].

MSE is the “mean of the square of the difference between actual and estimated values” [14]. RMSE is the “square root of the mean squared error (MSE)” [13]. For these metrics, the lower the value, the better the model. For R-squared, which is “the squared correlation between the observed outcome values and the predicted values by the model” [13], the higher the value the better the model.

References

[1] Drakos, Georgios. “Decision Tree Regressor Explained in Depth.” *GDCoder*, 5 Feb. 2020, gdcoder.com/decision-tree-regressor-explained-in-depth.

[2] “Decision Tree Regression.” *An Introduction to Data Science*, 2021, www.saedsayad.com/decision\_tree\_reg.htm#:%7E:text=Decision%20tree%20builds%20regression%20or,decision%20nodes%20and%20leaf%20nodes.

[3] ---. “Random Forest Regressor Explained in Depth.” *GDCoder*, 5 Feb. 2020, gdcoder.com/random-forest-regressor-explained-in-depth.

[4] Brownlee, Jason. “Linear Regression for Machine Learning.” *Machine Learning Mastery*, 15 Aug. 2020, machinelearningmastery.com/linear-regression-for-machine-learning.

[5] “Linear Regression — ML Glossary Documentation.” *Linear Regression*, 2017, ml-cheatsheet.readthedocs.io/en/latest/linear\_regression.html#:%7E:text=Linear%20Regression%20is%20a%20supervised,(e.g.%20cat%2C%20dog).

[6] Maklin, Cory. “Machine Learning Algorithms Part 11: Ridge Regression, Lasso Regression And Elastic-Net Regression.” *Medium*, 31 Dec. 2018, [medium.com/@corymaklin/machine-learning-algorithms-part-11-ridge-regression-7d5861c2bc76](mailto:medium.com/@corymaklin/machine-learning-algorithms-part-11-ridge-regression-7d5861c2bc76).

[7] “Difference between Ridge and Linear Regression.” *Data Science Stack Exchange*, 13 Mar. 2020, datascience.stackexchange.com/questions/69661/difference-between-ridge-and-linear-regression#:%7E:text=Linear%20Regression%20establishes%20a%20relationship,also%20known%20as%20regression%20line).&text=By%20adding%20a%20degree%20of,regression%20reduces%20the%20standard%20errors.

[8] Sethi, Alakh. “Support Vector Regression Tutorial for Machine Learning.” *Analytics Vidhya*, 1 Apr. 2020, www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning.

[9] Raj, Ashwin. “Unlocking the True Power of Support Vector Regression.” *Medium*, 5 Oct. 2020, towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:%7E:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.

[10] Raj, Ashwin. “Unlocking the True Power of Support Vector Regression.” *Medium*, 5 Oct. 2020, towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:%7E:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.

.

[11] Singh, Harshdeep. “Understanding Gradient Boosting Machines - Towards Data Science.” *Medium*, 4 Nov. 2018, towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[12] ---. “Strong Learners vs. Weak Learners in Ensemble Learning.” *Machine Learning Mastery*, 27 Apr. 2021, machinelearningmastery.com/strong-learners-vs-weak-learners-for-ensemble-learning/#:%7E:text=Weak%20learners%20are%20models%20that,boosting%20class%20of%20ensemble%20methods.

[13] “Regression Model Accuracy Metrics: R-Square, AIC, BIC, Cp and More.” *Articles - STHDA*, 11 Mar. 2018, www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more.

[14] Great Learning Team. “Mean Squared Error - Explained | What Is Mean Square Error?” *GreatLearning Blog: Free Resources What Matters to Shape Your Career!*, 27 Apr. 2021, [www.mygreatlearning.com/blog/mean-square-error-explained](http://www.mygreatlearning.com/blog/mean-square-error-explained).