Derek Graves

07/16/2024

Lab 4: Building a Regression Model

* Progress 1 - Screen shot pipeline ready to replace classifier with regression model

A screen shot of a computer

Description automatically generated

* Progress 2 - Screen shot performance of model on training and test sets.

A screenshot of a computer

Description automatically generated

* Progress 3 – Analysis: Compare performance and argue if model is underfitting the training set.
  + We can see that, in general, the training set metrics suggest that the model performs slightly better on the training data than on the test data, which is expected. Both the MAE, RMSE, and MSE scores are low, indicating that the predictions made by the model are close to the actual weights. The R² score only slightly drops from the training set (88.93%) to the test set (86.77%), indicating a good fit to the data on both sets. Overall, the close similarity between the performance metrics on both sets shows that the model can generalize well to unseen data.
  + The low error values and high R² scores suggest that the model is capturing the relationship between height and weight effectively and is not underfitting the training data.
* Progress 4 - Screen shot linear fit

A graph of a line with yellow and purple dots

Description automatically generated with medium confidence

* Submission 5 – Analysis
  + a) Will more training data improve the fit?
    - In my opinion, adding more training data would result in minimal improvement because our model is already performing well, and has minimal variance between the test set metrics, indicating it may be almost at peak performance already. Adding in more data could improve performance, but also might contribute to adding more noise than useful information, which may not be worth the trade-off.
  + b) Is the model under or over fitting?
    - Based on our performance metrics, such as the high R^2 scored for the training and test sets, we can conclude that the model doesn’t seem to be over or underfitting the data and that it generalizes well to unseen data. We also have low MAE and RMSE values showing that predictions are close to actual values. In my opinion, it seems like the model has a good balance overall, or else we would see the model performing well on either the test or training data, and poorly on the other, representing over-fitting or under-fitting.
* Submission 6 - Explain
  + Did adding age improve the model?
    - There was a slight increase in performance. RMSE on the set decreased from 5.31 to 5.18, and the R^2 score on the test set increased from 86.77% to 87.39%. There was a similarly small improvement for the test set.
  + Propose an explanation for the results.
    - The slight increase in performance shows that age does contribute slightly to increasing the predictive power of the model. While the improvement is not overly significant, the increase in the model's performance metrics across both training and test sets with age included shows that age provides additional insight into weight that height alone may not explain. This could be due to age-related changes in body composition and metabolism as one ages that affect weight beyond what can be explained with just height alone.
* Progress 7 - Scatter graph
  + Scatter graph with cubic model displayedA screen shot of a graph

    Description automatically generated
  + Propose an explanation
    - The cubic model seems to better capture the non-linear relationship between weight and height, especially where weight grows at a faster rate along with height. It seems to follow the clusters of data and adapt to the variations in the data where the linear model could not as easily.
* Submission 8 – Explain
  + Does the polynomial fit do better?
    - Yes, the polynomial fit did better. The RMSE and R^2 scores on the test set decreased from 5.31 to 4.31 and 86.77% to 91.30% respectively, indicating that the polynomial model seems to capture the relationship between height and weight better than the linear model.
  + Where does it fit the best?
    - It seems to fit best in the non-linear portions of the data, particularly where weight increases more rapidly with height during periods of more rapid growth. The polynomial fit seems to be able to better follow these variations in the data, where the linear model could not adapt and fit as well.

* Progress 9 - Screen shot of scatter plot with degree 8 polynomial fit

A screen shot of a graph

Description automatically generated

* Submission 10- Explain
  + a) Compare degree 3 and degree 8 fits
    - Overall, the degree 8 fit does marginally fit the data better than the degree 3 fit. Where the degree 3 fit was able to better follow clusters and variance in the data, the degree 8 fit seems to do so even better, but only leading to a marginal decrease in error rates and increase in R^2 values. For instance, the RMSE and R^2 values for the training set and test sets showed a very marginal increase. On the test set, the RMSE score only decreased by 0.05 (from 4.31 to 4.26) and the R^2 score was only improved by 0.20% (from 91.30% to 91.50%).
  + b) Is the increase in performance big enough to justify the degree 8 fit?
    - Given the additional complexity added by increasing our polynomial degree from 3 to 8, and the added potential for overfitting the data by doing so, the small increase in performance from the degree 8 fit may not be worth it. This could lead to higher training scores, but poor performance on unseen data. Since there isn’t a major advantage in improving metrics with the degree 8 fit, we may just want to stick with the simpler degree 3 fit to be safe.
* Progress 11 – Screen shot of scatter plot with degree 8 Elastic net.

A screen shot of a graph

Description automatically generated

* Submission 12 – Explain
  + Compare coefficients for regular and elastic net for degree 8.
    - There are several things that the differences in coefficients for regular and elastic net for degree 8 can tell us. The bias for example saw a significant reduction from 30.63 to 18.85, which shows the baseline might have been over-estimating. The degrees at quadratic and beyond seem to have been significantly reduced (the higher-degree terms in particular). This might indicate that the higher degree terms are not be adding a lot in predictive power to the model, and may just be adding additional complexity, which could lead to potential overfitting. The elastic net looks as though it is increasing the generalization of the model to avoid overfitting new, unseen test data. This goes to show that a simpler model may be the better option to capture this relationship.
  + Which ones have been reduced in elastic net?
    - The higher-degree coefficients (from quadratic and cubic to beyond) have been reduced. This shows that the higher-degree coefficients may not hold much weight in predictive power of the model, and so the elastic net model is showing that the simpler relationships may be more important than overly-complex degrees.