Name: Sanayya

Email address: sanayya1998@gmail.com

Contact number:

AnyDesk address:

Years of Work Experience:

Date: 06-02-2022

Self-Case Study -1: Don’t Overfit II: The Overfitting

Overview

1. One of the primary goals of predictive modelling is to create a model that can make accurate predictions based on previously unknown data. Checking that models haven't overfitted the training data, which can lead to suboptimal predictions on new data, is an essential stage in the modelling process.
2. Overfitting occurs when a model performs well with training data but fails miserably with test data. There are numerous reasons for overfitting, but we'd like to highlight a few of the most significant. The first is that training samples have fewer data points, the second is that the dataset is unbalanced, and the third, and most importantly, is that the model is complex.
3. We'll try out practically all of the machine learning strategies available to avoid overfitting in a dataset from a Kaggle competition called Don't Overfit II. This challenge aims to highlight existing algorithms, approaches, or tactics for avoiding overfitting.
4. The datasets consist of-
   - train.csv- This file consists of id, numerous continuous features whose meaning are not explicit and also the binary target which we have to predict.
   - test.csv- This file consists of id, and numerous continuous features whose meanings are not explicit.
5. The objective is given 250 data points and 300 columns in the training dataset; we have to create a model that accurately predicts the binary target for 19750 unknown data points in the test dataset without overfitting.
6. This is basically a classification problem. The metric for evaluation is provided as AUCROC between the predicted target and the actual target value. So, we will be working with AUCROC as a metric.

Research-Papers/Solutions/Architectures/Kernels

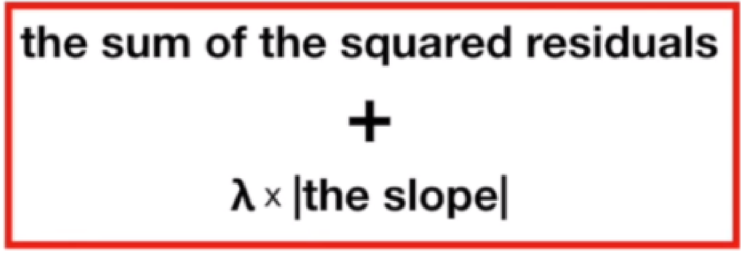
1. https://towardsdatascience.com/underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6fe4a8a49dbf
   a. Observations-
      - This blog has covered the strategies for avoiding overfitting in great depth and the justifications for using them.
      - The complexity of a model can be lowered by removing less important and irrelevant input (noise), which will help the model avoid overfitting and perform effectively on unknown data.
      - We should enlarge the data set since inadequate data causes the model to overfit so that it can avoid overfitting.
      - The model becomes increasingly sophisticated as the number of features increases. Adding a regularisation component to the cost function, rather than deleting some characteristics randomly, will reduce the effect of unnecessary features by decreasing their weights.

   b. Takeaways-
      - The importance of feature engineering cannot be underestimated. Features based on statistics for various columns must be developed.
      - To boost overall performance, different models' predictions must be combined.
2. https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624
   a. Observations-
      - This is the most voted blog for the "Don't Overfit II" Kaggle problem.
      - Machine Learning models such as Ridge, Lasso, ElasticNet, LassoLars, Bayesian Ridge regression, Logistic Regression, and SGD classifier were used to evaluate mean cross-validate score and standard deviation.
      - Logistic regression and stochastic gradient descent were the best-performing models with CV mean 0.7447916666666667 and 0.7333333333333333.
      - Grid Search and Random Search hyperparameter optimization techniques were applied on the selected models to improve their performances. Grid Search CV has improved the CV mean score to 0.789, and random search has improved the CV mean score to 0.780.

   b. Takeaways-
      - In order to avoid overfitting, we can apply hyperparameter optimization techniques.
3. https://www.kaggle.com/artgor/how-to-not-overfit
   a. Observations-
      - This is the most voted kernel in the "Don't Overfit II" competition.
      - After performing EDA on the features, it was observed that there were no highly correlated features that could be eliminated. Similarly, some basic modelling was performed, and it gave a CV mean score of 0.7226 for logistic regression.
      - According to ELI5, there were just 32 features that were relevant. Since the model had 34 non-zero features, so ELI5 deleted only 2. It gives an improvement from 0.7226 to 0.7486 on CV mean score.
      - Feature selection techniques such as permutation importance, SHAP, Mlextend Sequential Feature Selector were used to choose the essential features, but the CV mean score didn't improve much. Thus, feature selection isn't the best approach to handle overfitting.

    - Various models were compared, including Logistic Regression, Gaussian Naïve Bayes, AdaBoost Classifier, Extra Trees Classifier, Random Forest, Gaussian Process Classifier, Support Vector Classification, kNN, Bernoulli Naïve Bayes, SGD Classifier, and it was observed that Logistic Regression is superior to most models. Other models appear to be either overfit or unable to cope with this short dataset. It gives an improvement from 0.7486 to 0.831 on CV mean score.
    - Experiments were conducted using feature engineering techniques such as polynomial features, statistics, and distance features. Then, using the sklearn package, multiple features were selected, such as percentile, SelectKBest, and RFE, and applied Logistic Regression and GLM models. Even yet, the CV mean score remains below 0.80.
  b. Takeaways-
    - Logistic regression is the best-performing model for the given dataset.
    - Complex feature selection and feature engineering techniques aren't the best approach to handle overfitting for this dataset.

4. https://medium.com/analytics-vidhya/just-dont-overfit-e2fddd28eb29
  a. Observations-
    - This is one of the most incredible blogs I came across while working on this Kaggle problem.
    - This blog's primary strategy is to employ LASSOCV (Least Absolute Shrinkage and Selection Operator Cross-Validation). It consists of two terms: LASSO and CV (where CV is Cross-Validation).
    - To avoid overfitting, LASSO regression adds a penalty represented by a Greek letter symbol lambda multiplied by the slope of the line. It greatly helps in increasing bias and decreasing residual.



the sum of the squared residuals

+

λ × |the slope|

    - LASSO then uses cross-validation to calculate the output for different alpha values using the LASSOCV function.
    - The usage of this approach results in a Kaggle leader board score of 0.843.
  b. Takeaways-
    - Standardization or normalization should be performed to squeeze the data between 0 and 1.
    - As mentioned above, feature engineering is very important.

5. https://buckeye17.github.io/Dont-Overfit/
  a. Observations-
    - This study created fourteen models, six of which were ensembles based on eight base models. Only a minor improvement over the best base models was achieved using ensemble approaches.
    - The statistic ROC AUC was used to evaluate models because it is the metric used to score submissions. The ROC curves for the majority of the models were below.

b. Takeaways-

- As mentioned above, logistic regression is the best-performing model for the given dataset.

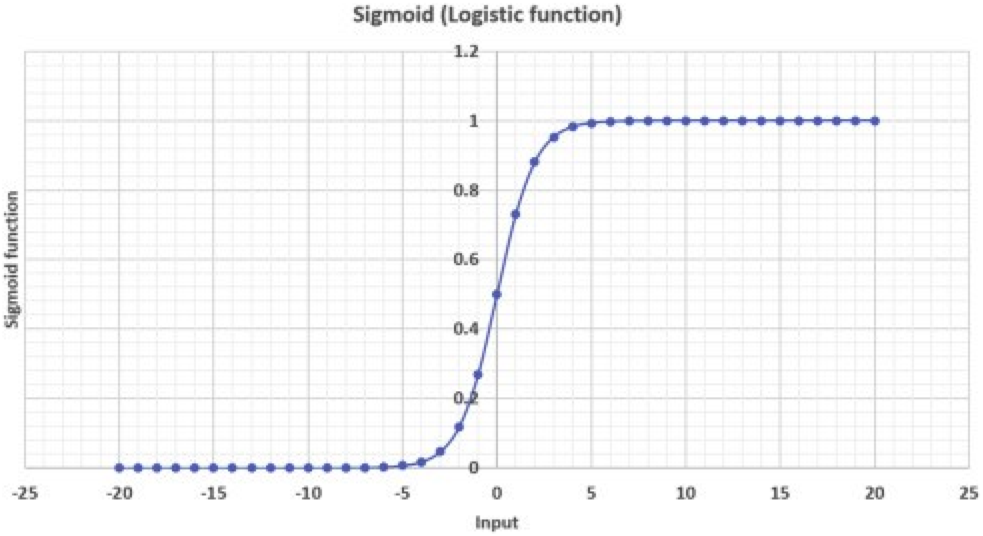
6. Logistic Regression

a. Understanding –

After reading blogs and solutions that are useful for solving this problem, we found out that Logistic Regression is one of the most common themes.
Logistic regression is a robust supervised ML algorithm used for binary classification problems (when the target is categorical). Despite its name, logistic regression is a classification model rather than a regression model. The logistic function defined below is used to model a binary output variable in logistic regression.

$$Logistic\ function = \frac{1}{1+e^{-x}}$$

where X is the input variable.
The main distinction between linear and logistic regression is that the range of logistic regression is limited to 0 and 1. Furthermore, logistic regression does not require a linear relationship between input and output variables, unlike linear regression. This is due to the odds ratio being transformed via a nonlinear log transformation.



For example, let's input the logistic function values ranging from -20 to 20. The inputs have been changed to a range of 0 to 1, as seen in the above figure.
The loss function used in logistic regression is called "maximum likelihood estimation (MLE)," and it is a conditional probability. The predictions will be categorized as class 0 if the probability is more prominent than 0.5. Otherwise, class 1 is allocated.

b. References-

- https://www.sciencedirect.com/science/article/pii/B9780128219294000044
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

First Cut Approach

Based on the research and readings that we have done. We will follow the below steps-

1. Using EDA, we'll build summary statistics for the dataset's numerical data and construct multiple graphical representations to understand the data better.
2. Feature engineering is one of the most critical aspects of this Kaggle problem. So we will look to come up with various new statistics features, including basic mathematical operations, trigonometric functions, hyperbolic functions, exponents and logarithms, and polynomial operations.
3. Since a large number of features could lead to overfitting, we will choose only the most essential features by applying PCA and Truncated SVD.
4. As the dataset is slightly imbalanced, we can use both oversampling and undersampling techniques to balance it.
5. We’ll apply various machine learning models, including kNN, Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Trees with hyperparameter tuning. We’ll also try to build even more powerful models using ensembling, keeping Logistic Regression as the base model.