Modelling201 Assignment - Khushal Soni

1. Types of variables –

Nominal, Ordinal, Interval and Ratio.

* Nominal- When we categories a variable in two or more categories but they are not in order. For e.g. – Sex of a person (Male or Female), race (black, Hispanic, oriental, white, other), political party (democrat, republican, other), blood type (A, B, AB, O) etc.
* Ordinal – Categorizing a variable where categories are in order. So, the things being measured are in some order. E.g. - football top 5 teams, pop music top 5 songs.
* Interval - An interval variable is a one where the difference between two values is meaningful. We can calculate mean, median, mode and we can perform addition, subtraction but division and multiplication is not applicable. There is no true zero.

e.g. – Temperature, SAT Score, IQ Score.

* Ratio - A ratio variable, has all the properties of an interval variable, but also has a clear definition of 0. For e.g. - mass, weight, and volume, distance in metre etc. Here, division, multiplication are applicable.

1. Difference between Multiple Linear Regression and Multiplicative Linear Regression –

* Multiple linear regression is a statistical technique that employs more than one independent variable to forecast the dependent variable. To illustrate, when attempting to estimate a person's wage based on their age, education, and work experience, multiple linear regression can be employed. In this scenario, the dependent variable is the salary, while the independent variables are age, education, and work experience.
* In multiplicative regression, the dependent variable is modelled as a function of the product of the independent variables. Multiplicative regression is particularly useful when analysing data that involves nonlinear relationships or interactions between variables that cannot be adequately captured by linear models.

1. Pros and Cons of Logistic Regression.

Pros-

* Can be used with both continuous and categorical independent variable.
* It is robust to outliers.

Cons-

* Logistic regression assumes that the independent variables are not highly correlated with each other (i.e., multicollinearity), as this can lead to unstable estimates of the regression coefficients.
* Logistic regression assumes that the observations are independent of each other (i.e., no autocorrelation), which may not be true in some datasets.
* Logistic regression assumes that there is a linear relationship between the independent variables and the natural logarithm of the odds of the dependent variable. However, this assumption may not always hold true in all cases.

1. Explain bias and Variance with Real world examples.

* Bias refers to the difference between the expected (or average) prediction of a model and the true value of the target variable or we can say when Train data is not a very good sample of actual data. A model with high bias underfits the data, as it fails to capture the complex patterns in the data and makes oversimplified assumptions.
* In simple terms when our model is trying to pass through every train set data points such that accuracy is very high, we can get high error for test set and our prediction on new data set will highly diverge from actual scenario. In technical words variance is defined as the degree to which the predictions of a model would vary if it were trained on a different dataset. A model with high variance tends to overfit the training data by capturing the noise and random fluctuations, resulting in poor performance on new, unseen data. This is because the model is too complex and cannot generalize well to new data.
* Example- Suppose we have a dataset of patients with different symptoms and corresponding diagnosis (e.g., cancer or not cancer). If we fit a simple model such as a linear regression or logistic regression to the data, the model may have high bias as it cannot capture the complex relationship between the symptoms and the diagnosis. This can result in a model that underfits the data and fails to identify the underlying patterns in the data. As a result, the model may fail to accurately diagnose patients with cancer, leading to false negatives or false positives. On the other hand, if we fit a very complex model such as a decision tree the model may have high variance as it can adapt too closely to the variability in the training data. This can result in a model that performs well on the training data, but poorly on new, unseen data. As a result, the model may overfit the data and may lead to overdiagnosis or underdiagnosis of patients.
* Example of bias- Suppose a company wants to know if a new training program improved their employee's performance. They ask the employees who completed the program about their experience and how it impacted their work. However, if they only ask those who voluntarily enrolled in the program and not those who did not. This could be biased because those who enrolled may have been more motivated or skilled to begin with, making them more likely to report positive results. This bias can lead to incorrect conclusions about the effectiveness of the program.
* Example of variance - Suppose we have a machine learning model that predicts the temperature for the next day based on various weather variables. If we train the model on a dataset that has many variations in weather patterns, the model may be too complicated and have high variance. This means that the model will perform well on the training data but may not perform well on new data. For example, if we train the model using data from a region with a specific weather pattern, such as a desert, it may not work well on data from a region with a more variable weather pattern, such as a coastal region. This is because the model has become too closely adapted to the training data and the unique characteristics of that region, leading to poor generalization to new data. Hence, the model's predictions may be unreliable and inaccurate in these areas with more variable weather patterns, which could have severe consequences such as inaccurate weather forecasts.
* Top of Form

1. What are the splitting Criterions for ID3: Iterative Dichotomiser, C4.5, CART: Classification and Regression Tree, CHAID: Chi-square automatic interaction detection, MARS: multivariate adaptive regression splines.

* ID3: Iterative Dichotomiser uses the entropy to select the best feature to split data.
* C4.5 uses information gain and feature with highest information gain is made root node.
* CART: Classification and Regression Tree uses Gini index as attribute selection measure.
* CHAID: Chi-square automatic interaction detection uses Chi-Square as a attribute selection measure.
* MARS: multivariate adaptive regression splines – uses forward selection and backward elimination attribute selection measure.

Forward selection: MARS starts with an empty model and adds basis functions (e.g., linear, hinge, or spline) one by one until the desired model complexity is achieved. At each step, the attribute that yields the largest improvement in the residual sum of squares (RSS) is selected as the best attribute for adding a basis function.

Backward elimination: MARS then uses backward elimination to remove the basic functions that contribute least to the model. At each step, the basis function that yields the smallest improvement in the RSS is removed.

1. What do you mean by giving high weightage in boosting?

* It typically refers to assigning higher weights to the misclassified samples in each round of boosting to help the algorithm focus on these samples and improve the classification accuracy.
* By giving higher weights to the misclassified samples, boosting aims to correct the errors made by the previous weak learners and improve the overall accuracy of the model.

1. What Is difference between AdaBoost, Gradient Boost and XGBoost?

* Adaboost is a boosting algorithm that aims to enhance the performance of a model by modifying the weights of samples that are misclassified. It achieves this by assigning greater weightage to misclassified samples and training the model on these weighted samples in order to minimize the error rate.
* Gradient Boost model uses gradient descent algorithm to minimize the loss function such as mean squared error the previous model. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.
* During each iteration in Adaboost the weights of misclassified samples are increases so that next weak learner focuses more on these samples. While Gradient Boosting updates the weights by computing the negative gradient of the loss function with respect to the predicted output. Adaboost uses decision stumps as weak learners while larger trees are allowed in gradient boost.
* Xgboost uses regularization techniques hence it is regularized form of existing gradient boost algorithm.
* Xgboost algorithm can handle missing values while other two can’t.

1. Example of stacking Ensemble method.

To predict whether a customer will cancel their subscription from a subscription-based service, we can use the stacking ensemble method to combine predictions from multiple base models such as Logistic Regression, Decision Trees. To do this, we first split the data into two sets: one for training the base models (training set A) and the other for training the meta-model (training set B). The base models are trained on training set A and generate predictions for the test data. The predicted values generated by the base models are then used as input features for the meta-model, which is trained on training set B using the actual churn status as the target variable. Finally, the trained meta-model is used to predict the churn status of the test customers based on the predictions generated by the base models. The final prediction is the average or weighted average of the predictions made by the meta-model. By combining the strengths of multiple models, the stacking approach has the potential to improve the performance of the model.

1. Reinforcement learning.

* Reinforcement learning is a type of machine learning where an agent learns to make decisions through trial-and-error interactions with an environment. The agent receives rewards or penalties for its actions and adjusts its behaviour to maximize its cumulative reward over time. It is similar to how humans learn from experience and feedback to make better decisions in the future.