**Ensemble Techniques Exam:**

Q1. What is an Ensemble Model?

An Ensemble model is a model formed by where a number of weak learners (classifiers/regressors that are barely better than guessing) combine (through averaging or max vote) to create a strong learner that can make accurate predictions of predictive analytics and data mining applications.

Q2. What is the difference between Bagging and Boosting. Explain it using an example?

|  |  |
| --- | --- |
| Bagging | Boosting |
| Bagging means that you take bootstrap samples (with replacement) of your data set and each sample trains a (potentially) weak learner. | Boosting, on the other hand, uses all data to train each learner, but instances that were misclassified by the previous learners are given more weight so that subsequent learners give more focus to them during training. |
| In bagging, different models run in parallel. | In boosting, the models run in a sequential manner where in each iteration more weight is given to misclassified points. |
| It creates an aggregated model with less variance | Bias is reduced |
| Overfitting is less, hence model may tend to underfit | Model may tend to overfit |

Q3 Explain the concept of Bias Variance Trade Off?

Whenever we build model prediction, it’s important to know prediction errors -bias and variance. There is a trade-off between a model’s ability to minimize bias and variance. Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias gives very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data. In supervised learning, **underfitting** happens when a model is unable to capture the underlying pattern of the data. Such models usually have high bias and low variance. On the other hand, Variance is the variability of model prediction for a given data point or a value which gives us the spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before. As a result, such models perform very well on training data but not on test data.In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

Q4 Inference for Car data set

our model we should build will be using Random Forest Regressor with XG Boost Regressor as we get the least RMSE value in this case that is 0.39

Q5. Inference for Prima Diabetes Data set

The max accuracy we are getting is by Logistic Regression and applying XG boost on it, is 79.65 %

So, for this data set our model will use Logistic Regression and XGBoost ensemble technique for maximum Likelihood.

Q6 Inference for Titanic Dataset

Our Model has give the accuracy of 84.32% with all the models in account. Which means it correctly predicts 84.32 % accuracy if a person survived or not