**Bagging on Low Variance Models.**

Bagging (also known as bootstrap aggregation) is a technique in which we take multiple samples repeatedly with replacement according to uniform probability distribution and fit a model on it. It combines multiple predictions to give a better prediction by majority vote or taking the aggregate of the predictions. This technique is effective on models which tend to overfit on the dataset (high variance models) like. Bagging reduces the variance without making the predictions biased. This technique acts as a base to many ensemble techniques so understanding the intuition behind it is crucial.

If this technique is so good, why do we use it only on models which show high variance? What happens when we use it with models which have low variance? Let us try to understand the underlying issue with the help of a demonstration.

I will be using bagging on decision tree to prove that bagging improves the accuracy for high variance models and compare it to bagging on simple linear regression which is biased depending on the dataset. Simple linear regression is biased when the predictor is not perfectly correlated to the target variable.

**Bias and Variance**

We will be talking about Bias and Variance throughout the article so let us get an idea of what it is first.

High bias refers to the oversimplification of the model. i.e. When we are unable to capture the true relation of the training dataset. Our objective of creating a model is to capture the true nature of the dataset which makes high bias an undesired phenomenon.

High variance refers to the situation when we are overcomplicating our model. i.e. The situation where, in the process of capturing the true nature of the model, we are creating a model which learns the training data so well that its accuracy deteriorates on any other dataset. This situation is also undesired as our objective is to make predictions for unseen data.

When we are creating a model, we want to strike a balance between the bias and variance. Bias and Variance are opposite of each other so whenever we try to reduce the variance, we are increasing the bias of the model at the same time. This dilemma of overfitting/underfitting is called Bias-Variance Tradeoff. This image gives a good idea of their relation witch each other.

**High Variance Model – Decision Tree**

Decision Tree tries to classify the target variable and with default settings, it does not stop unless it classifies every category perfectly. This makes the tree overfit on the data provided and the accuracy of the model on test dataset will be low. Let us verify this using a dataset. We will use Pima Indians Diabetes Dataset. <Insert dataset hyperlink>

Accuracy for the Decision Tree:

<Insert code for decision tree here>



**Bagging of Decision Tree**

As we have discussed earlier, bagging should decrease the variance in our predictions without increasing the bias. The direct effect of this property can be seen on the change in accuracy of the predictions. Bagging will make the difference between training accuracy and test accuracy smaller. We might not always observe a change in the training accuracy, but the test accuracy will always be better in this case. Let us check what happens when we use it on our dataset.

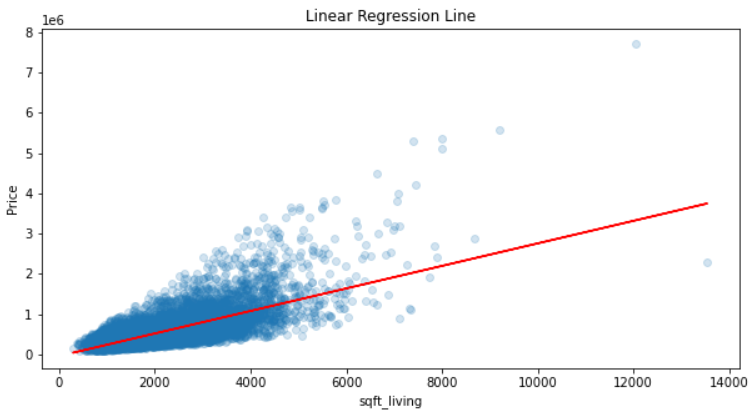
<Insert code for bagging on decision tree here>



This result proves our point! Bagging does work on high variance models!

**Low Variance Model – Simple Linear Regression**

Everything we talked about till now in this article is known and intuitive but what happens when we try to use Bagging on a low variance model like Simple Linear Regression? Intuitively, we may assume that when the variance of the model is already low and bagging does not increase the bias, bagging will not be able to reduce it further and as a result the accuracy of the model will not change. **Wrong!!**

To understand what exactly happens in this scenario, I will use King County House Pricing Dataset <Link to the dataset used>. I am using only one variable to make the visualization easier. Let us look at the scatter plot between price and sqft\_living to get a general idea of the relation between the two and build a simple linear regression model on it.

Accuracy of simple linear regression model in this case:

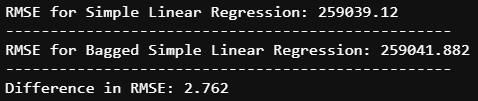
<Code for fitting the data as well accuracy calculation>



**Bagging on Simple Linear Regression**

Before discussing ‘the curious case of Bagging on Simple Linear Regression’, let us quickly check whether the accuracy after bagging simple linear regressions improves or even improves the accuracy of the model.

<Code for fitting and predicting on simple linear regression as well as bagging on simple linear regression>

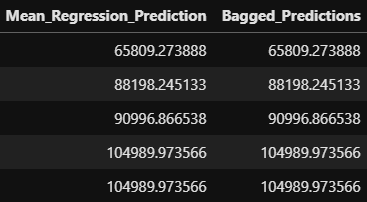


Despite bagging 200 simple linear regression models, the RMSE (Root Mean Squared Error) went down by 2.7 points! The reason behind this is that simple linear regression model is biased. We observe a very small change in the RMSE because our model was able to capture the trend of the dataset quite well. The correlation between sqft\_living and price was 0.7 and simple linear regression model captures the linear relation between the variables. If the correlation for simple linear regression was lower, the difference in RMSE would have been larger. We will discuss why in the next section.

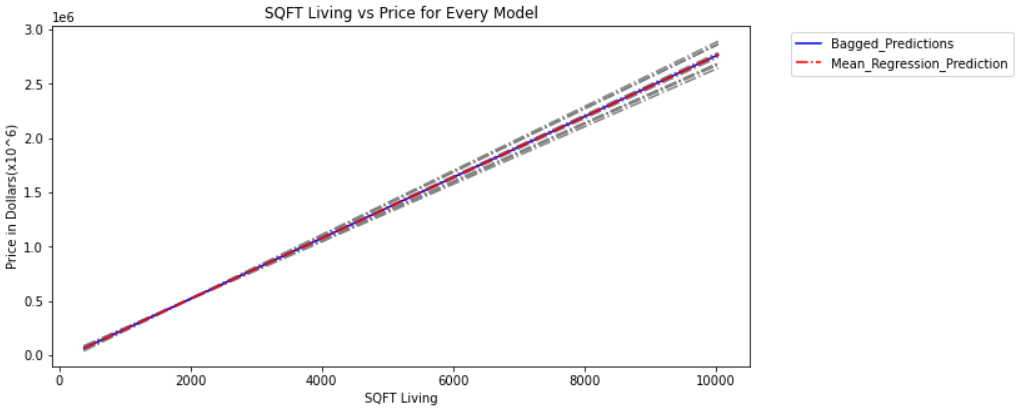
**Why doesn’t Simple Linear Regression work well with Bagging?**

Simple Linear Regression is a conditionally biased model. i.e. when there is clear linear relationship between the variables, this model can be considered stable. In this scenario, the accuracy of the model will not decrease even if we use bagging on it. Even when this relationship between the variables changes, simple linear regression tries to create a straight line to capture the trend between the data. In this scenario, the model becomes biased.

As the correlation between our variables are 0.7, it is not perfectly correlated but it is good enough that the use of simple linear regression makes sense here. This is the reason because of which the RMSE did not take a huge dip when we used bagging on it. The bagging technique creates multiple linear regression model and takes the mean of their predictions. All these points will lie on the regression line which can be produced by taking the mean of intercepts and coefficients of every model. To prove this point, let us compare the predictions obtained from the bagged model and the one which we would get if we generate a line which we just talked about.



Every model which was created during the bagging process was more biased than the simple linear regression model which we could have obtained by using the entire dataset. Let us look at what the biased regression lines in comparison to the final bagged predictions.



Every grey line present in this graph represents individual simple linear regression models. Their bias is higher than the regression line which could have been obtained from the original dataset directly as many duplicate points are present in bootstrap samples. This gives some points more leverage than the rest.

**Conclusion**

References:

[1] <https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/>  
[2] Daniel T. Larose, Chantal D. Larose, Data Mining and Predictive Analytics  
[3] King County House Pricing - <https://www.kaggle.com/swathiachath/kc-housesales-data>  
[4] Pima Indians Diabetes - <https://www.kaggle.com/uciml/pima-indians-diabetes-database>  
[5]   
[6]   
[7]