**Baseline model for US Tuition Prediction**

**Sunny Kumar Kandula**

**<subject>**

**<University>**

**<Professor>**

***Introduction***

As I mentioned in my previous proposal this project aims to provide a machine learning model that can predict the total cost of fee that could be incurred by a student depending upon their choice of study.

For this deliverable the goal is provide an baseline model, as the definition suggests, a baseline model is a model that does not require much technical understanding and can be easily implemented to draw some conclusion and results/predictions and metrics, which could be used as metrics that can be weighed against other models in order to make them perform better than our reference model.

In order to train my baseline models I had to encode my data, on the whole this dataset contains 5 features and 1 target variable that has to be predicted. And out of these 5 features, 4 features were categorical in nature and only 1 was continuous. Hence there was a need to perform OneHotEncoding for this data to be fitted to a Regression model. Hence I decided to use sklearn.preprocessing technique/method called OneHotEncoder which automatically performs the necessary steps.

***Baseline Median Model***

Upon looking at ‘y’ target variable we can understand that it contains high variance! The baseline median model uses the median value ***10,286*** from the training values of y, this gives us our median which can be used to make predictions.

The predictions were made using this model, and we get the following results:

* Mean Absolute Error (MAE): 5,821.28
* Mean Squared Error (MSE): 82,693,049.63
* Root Mean Squared Error (RMSE): 9,093.57

These results are very large in numbers which defines the fact that our model is not capturing the data well. However these metrics are just based on the metrics yielded by our baseline Median model.

***Mean Baseline Model***

A mean baseline model aims to predict mean value of the target variable, and just like the median baseline model it serves as a reference to compare our predictions.

For this I calculated the mean of the target variable, ‘value’. And then used this to predict rest of the values. And tried to evaluate this model. And I got these results:

* Mean Absolute Error (MAE): 5,821.28
* Mean Squared Error (MSE): 82,693,049.63
* Root Mean Squared Error (RMSE): 9,093.57

As we can see the metrics are same for both these models. In this case both models are performing the same way!

***Linear Regression***

My primary baseline model that I shall use as a reference against other models which I hope will perform way better than Linear Regression, the reasons for choosing Linear Regression is the fact that it is one of the simplest Regression models which performs the task of predicting values of a variable which is dependent on the other.

Linear Regression is easy to implement and doesn’t take much time to train as well. Linear Regression gives us idea of relationships between features. And most importantly its easy to compare results against other complex models.

Linear Regression is not always the best choice, for a regression task, the model that we choose depends on the dataset we have and the task at hand. How ever once trained on the training set and testing it against some testset. We yield the following results for a simple Linear Regression Model:

* Mean Squared Error (MSE): 23,649,223.47
* R-squared (r2\_score): 0.687

Lets compare these against the other baseline models. Whose MSE is 82,693,049. Upon comparing these values we notice a significantly lower MSE score for the Linear Regression model with an r2\_score of 0.68 which is an average score. This could be due to lack of larger feature space and more number of categorical features. This is exactly why I would use a DecisionTreeRegressor as my primary model.

The obtained r2\_score is a good start. But there is a lot of variance in each categorical feature and this is not being captured properly within our Linear Regression model.

In conclusion Linear Regression model outperforms the Baseline Median and Baseline Mean models and Linear Regression shall act as my reference model when it has to be compared with other models.

For on coming deliverables I intend to train RandomForestRegressor and DecisionTreeRegressor as I feel these would be a better fit for the data I have. I hope to perform some feature engineering to capture some additional insights which could possibly help in better predictions along with hyper parameter tuning in order to find the optimal hyper parameter.