TANZANIA TOURISM PREDICTION

The objective of this hackathon is to develop a machine learning model to predict what a tourist will spend when visiting Tanzania.

DATA UNDERSTANDING

Column Name	Definition
id	Unique identifier for each tourist
country	The country a tourist coming from.
age_group	The age group of a tourist.
travel_with	The relation of people a tourist travel with to Tanzania
total_female	Total number of females
total_male	Total number of males
purpose	The purpose of visiting Tanzania
main_activity	The main activity of tourism in Tanzania
infor_source	The source of information about tourism in Tanzania
tour_arrangment	The arrangment of visiting Tanzania
package_transport_int	If the tour package include international transportation service
package_accomodation	If the tour package include accommodation service
package_food	If the tour package include food service
package_transport_tz	If the tour package include transport service within Tanzania
package_sightseeing	If the tour package include sightseeing service
package_guided_tour	If the tour package include tour guide
package_insurance	if the tour package include insurance service
night_mainland	Number of nights a tourist spent in Tanzania mainland
night_zanzibar	Number of nights a tourist spent in Zanzibar
payment_mode	The mode of payment for tourism service
first_trip_tz	If it was a first trip to Tanzania
most_impressing	what impressed a toursit in Tanzania
total_cost	The total tourist expenditure in TZS(currency)

E.D.A

```
ID                          0
country                     0
age_group                   0
travel_with              1114
total_female                3
total_male                  5
purpose                     0
main_activity               0
info_source                 0
tour_arrangement            0
package_transport_int       0
package_accomodation        0
package_food                0
package_transport_tz        0
package_sightseeing         0
package_guided_tour         0
package_insurance           0
night_mainland              0
night_zanzibar              0
payment_mode                0
first_trip_tz               0
most_impressing           313
total_cost                  0
dtype: int64
```

COLUMNS WITH MISSING DATA

Most missing values

1. Travel_with
2. Most_impressing

I dropped these columns

Few missing values

3. Total_male
4. Total_female

Replaced with mean

E.D.A

- **<u>COLUMNS DROPPED</u>**
- **Most_impressing**
- **Travel_with**
- **ID**
- **Combined total_male with total_female to give me total tourists then I dropped them(due to high multicollinearity)**
- **Info_source**

E.D.A

```
age_group                0
total_female             0
total_male               0
purpose                  0
main_activity            0
info_source              0
tour_arrangement         0
package_transport_int    0
package_accomodation     0
package_food             0
package_transport_tz     0
package_sightseeing      0
package_guided_tour      0
package_insurance        0
night_mainland           0
night_zanzibar           0
payment_mode             0
first_trip_tz            0
total_cost               0
dtype: int64
```

COLUMNS I WAS LEFT WITH

DATA MODELING

- Encoded the categorical using one-hot encoding
- Scaled the numerical data
- Split it into train and test sets(25%)

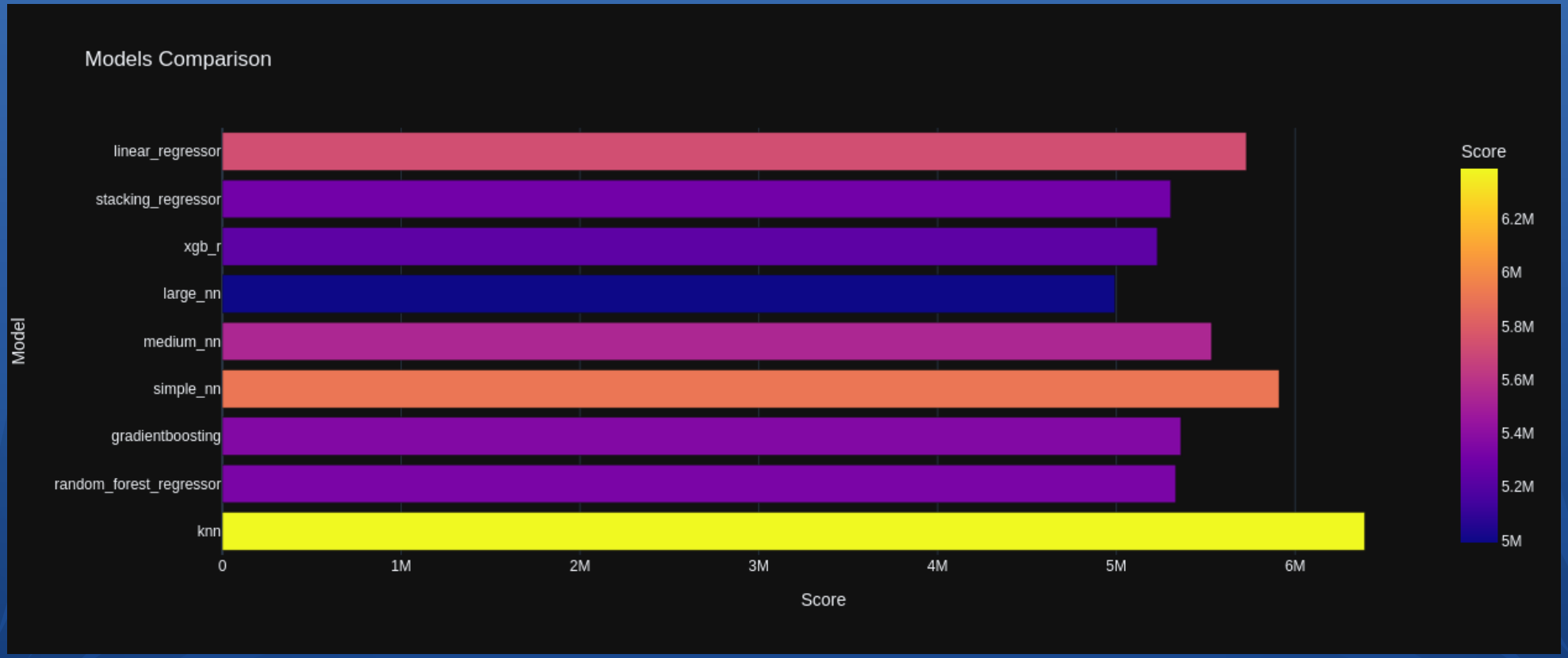
MODEL BUILDING AND EVALUATION

- K-nearest-neighbour
- Linear regression
- Random forest regressor
- Gradient boosting
- Stacking regressors
- Neural network
- Xgb_boost

MODEL EVALUATION

	Model	Score
5	large_nn	4.992640e+06
6	xgb_r	5.229569e+06
7	stacking_regressor	5.304558e+06
1	random_forest_regressor	5.332560e+06
2	gradientboosting	5.361186e+06
4	medium_nn	5.533231e+06
8	linear_regressor	5.727607e+06
3	simple_nn	5.909512e+06
0	knn	6.388657e+06

MODEL EVALUATION



NEURAL NETWORK

```
large_nn = Sequential()
large_nn.add(InputLayer((32,)))
large_nn.add(Dense(512, 'relu'))
large_nn.add(Dropout(0.01))
large_nn.add(Dense(256, 'relu'))
large_nn.add(Dropout(0.01))
large_nn.add(Dense(128, 'relu'))
large_nn.add(Dropout(0.01))
large_nn.add(Dense(64, 'relu'))
large_nn.add(Dropout(0.01))
large_nn.add(Dense(32, 'relu'))
large_nn.add(Dropout(0.01))
large_nn.add(Dense(1, 'linear'))

opt = Adam(learning_rate=.1)
cp = ModelCheckpoint('/content/drive/MyDrive/Colab Notebooks/models/large_nn4', save_best_only=True)
large_nn.compile(optimizer=opt, loss='mse', metrics=[MeanAbsoluteError()])
large_nn.fit(x=X_train, y=y_train, validation_data=(X_test, y_test), callbacks=[cp], epochs=100)
```

ZINDI SCORE

- PUBLIC LEADERBOARD SCORE : 5295675.062