1. Tuning and improving machine learning model

In this part I made some changes:

* I use direct zipcode instead of shortened zipcode
* Add two more features ‘agi’ and ‘N1’ to cleaned\_taxes table
* I used ‘inner join’ instead of ‘outer’ join so that we do not have to fill NA with some random value.
* Reduce number of trees so that we can train model faster.
* All metrics improve compare with the original model.

Detailed codes are found in following link:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3514859587127367/1338417224023826/5495093447822595/latest.html>

1. Build models to predict diamond price

In this part, I experimented with three different models:

* Linear Regression Model
* Random Forest
* Gradient Boosting

Linear Regression model perform worst among three models. Gradient Boosting model slightly outperforms Random Forest; both models achieve much stronger results comparing with Linear Model. Detailed codes are found here:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3514859587127367/202736143104595/5495093447822595/latest.html>