EVALUATION MODEL REPORT

I. Visualize data

There are four csv data file which contain both ground truth and the prediction from three different models. Each of them contain two columns – “order_id” and “final_price” - each column hold 315 variables.

Visualizing ground truth data:

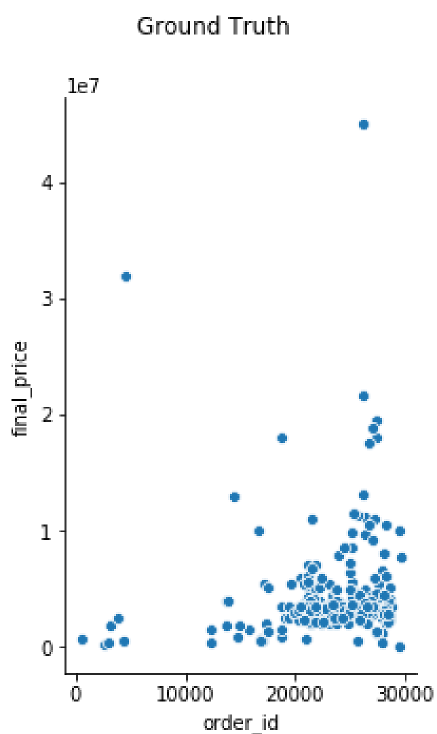

Figure 1. Visualize ground truth

Then visualizing prediction results:

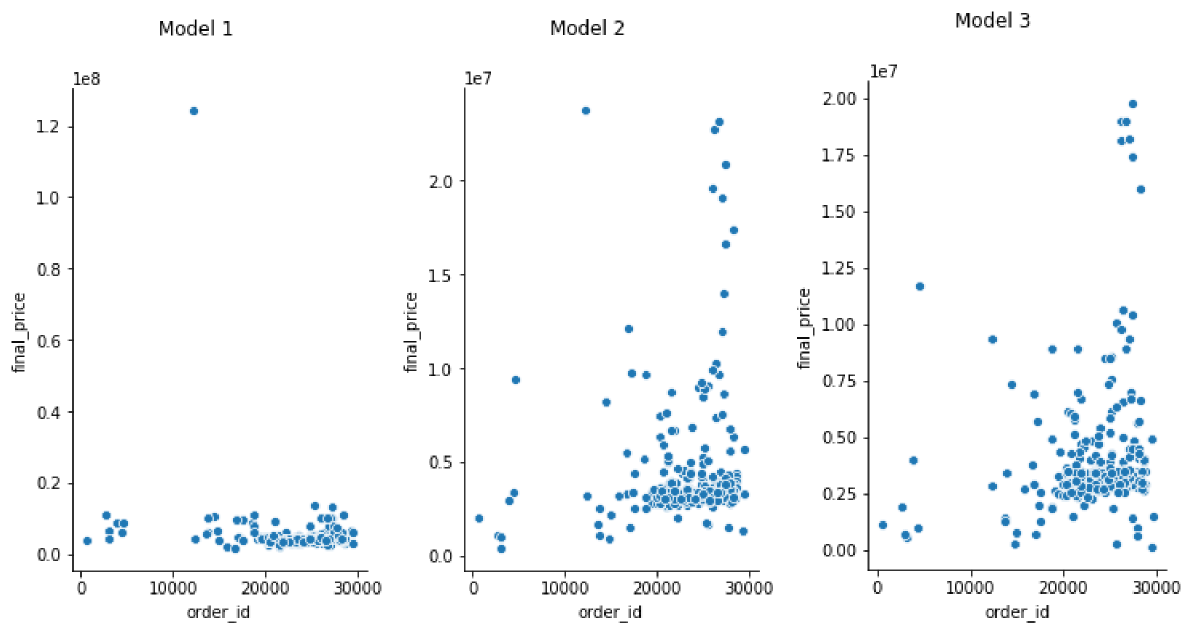

Figure 2. Visualize prediction results

By visualizing data, we can see that the “final_price” in ground truth get the mean between 0 and 1 with the variance in that range too.

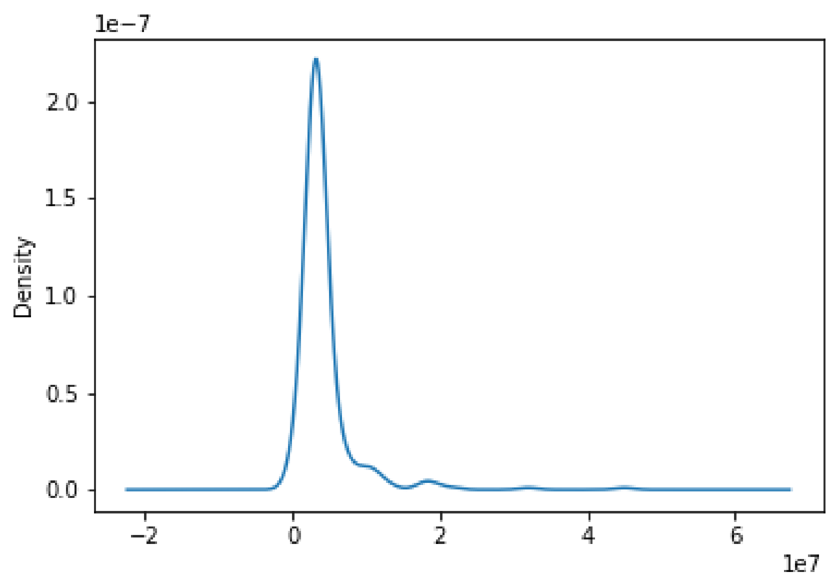

Figure 3. Density chart of ground truth

The first model results just stand at range 0 to 0.2, so the error between it and ground truth.

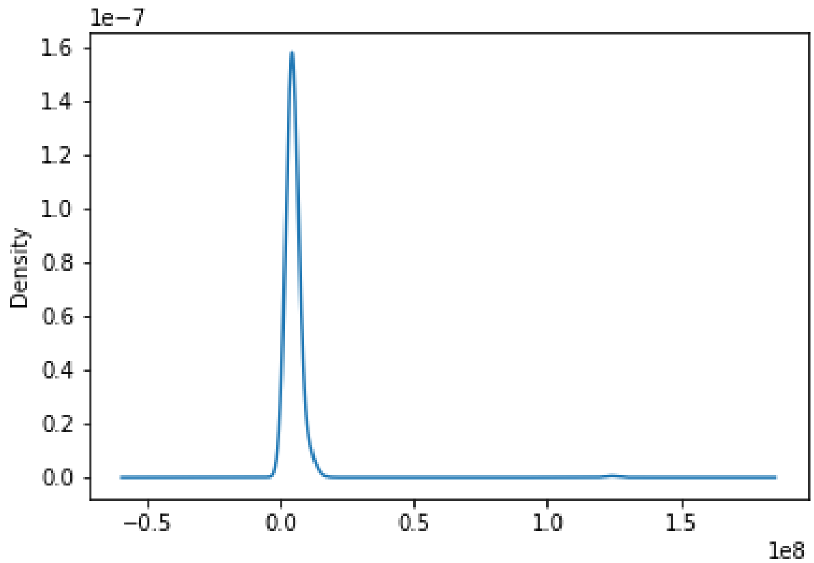

Figure 4. Density chart of model 1

The second and the third are much better, their results are scattered around range 0 to one, and get some variable at higher range. Their density chart look more like ground truth data.

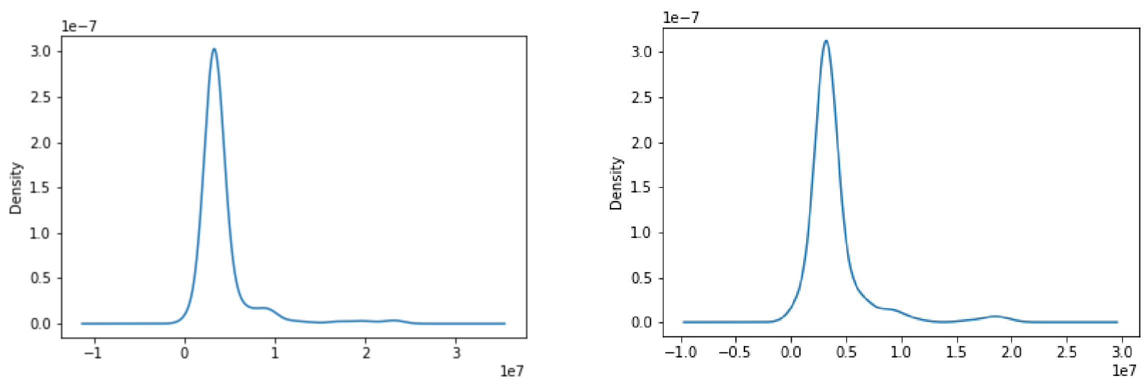

Figure 5. Density char of model 2 and 3

II. Validate Logivan's models

When choose metrics to validate these models. I confused between Mean Absolute Error (MAE) and Root mean squared error (RMSE) because they are two most common used to measure accuracy for continuous variables.

Take a look on their equations:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

With the meaning of parameters are:

- n **:** The number of samples that we calculated
- y_j : The prediction result from model with the sample number is j
- $\hat{y}_j$: The truth result of that sample.

Both of them express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.

However, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable. Therefore, I decided to choose RMSE to validate these model.

Follow the formula, I calculated the **RMSE** of these models:

- Model 1: 7898391
- Model 2: 2886121
- Model 3: 2460703

III. Conclusion and discussion

The **RMSE** results fit with the prediction in first part:

- Model 1 is the worst model is the one with the highest **RMSE**.
- Model 3 is slightly better than model 2 but it is acceptable.

Base on the visualized and **RMSE** results, assume that these data belong to the test dataset which mean the model is built without using the ground truth data. I think the model 1 get underfitting problem, there are two main reason of that problem:

- Too few data.
- The model is too simple.

However, model two and three performance are well, so we can reject the first reason. I think we should add more polynomial variable for this model, or even use machine learning to get the better result.

The best model base on the **RMSE,** it should be model 3. With 315 number of samples, 2460703 is a good **RMSE** result for an regression problem.