**Tell us how you validate your model, which, and why you chose such evaluation technique(s).**

1. Since the Kaggle dataset is the test dataset, I split the training dataset into a train set and a validation set to evaluate the performance of the model after we have obtained the best training model.
2. I used a 5-fold cross validation on the training set to prevent overfitting since we are evaluating it over the average result. To tune the model, I used GridSearchCV to iterate through a few possible values for key parameters to obtain the best possible mode.
3. To evaluate the model, I fit the best model from step 2 onto the validation dataset and obtained the ROC AUC score to evaluate the performance since it is the metric used to evaluate the models for this particular Kaggle competition.

**What is AUC? Why do you think AUC was used as the evaluation metric for such a problem? What are other metrics that you think would also be suitable for this competition?**

AUC refers to the Aurea under the Curve

What insight(s) do you have from your model? What is your preliminary analysis of the given dataset?

Based on the final model and the rest of the models

**Can you get into the top 100 of the private leaderboard, or even higher?**

I did not manage to but it was close and would likely be possible if I had more time (and daily tries on kaggle) to test out other models such as XGBoost or to further finetune the LightGBM model with RandomizedSearchCV. It would also be good if we could ensemble different models together, although that might affect the interpretability.

Scores

Private: 0.86619 (Rank 166 if competition was still open), just 0.00104 less than 100th place

Public: 0.86034

Graphical user interface, application

Description automatically generated