Final Report

04/25/2021

Team Outliers

- Jared Mlekush (kaggle id: jaredmlekush)
- Berkay Canogullari (kaggle id: berkaycanogullari1)

Abstract

Performed in depth analysis of data that describes miRNA features with a target cancer stage. Followed with EDA steps to decide how to structure our model. After doing the necessary analyses we implemented XGBoost and LightGBM, which perform particularly well on tabular data. Afterwards, we used the Optuna framework to optimize the hyperparameters of the selected model. Then we presented the associated results for both the base model and hyperparameter optimized model.

Dataset

Description:

Data consists of 8302 observations with 979 features consisting of miRNA data. From pandas description feature, we found/saw that the data was of type float and integer. The data given was clean - There was no missing data, it was already Min Max scaled, and anonymized, which led to no necessary feature engineering. Using this, our goal is to use Machine learning to analyze the features and predict the stage of cancer the patient has.

EDA:

First we analyzed the target distribution to figure whether we are dealing with an imbalanced or balanced classification problem.

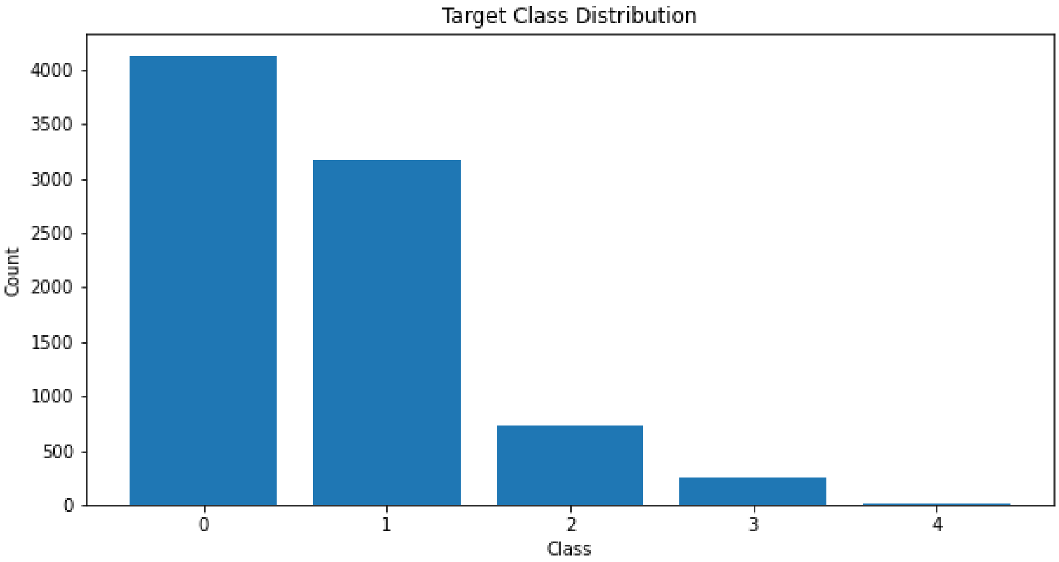

Figure 1

As you can see from Figure 1 we concluded that we are dealing with an imbalanced classification problem. Because of this imbalance, we wanted to see if there was an association between this imbalance and the only categorical variable "Problem_id". First we analyzed problem_id distribution. Then we checked class distribution per problem_id.

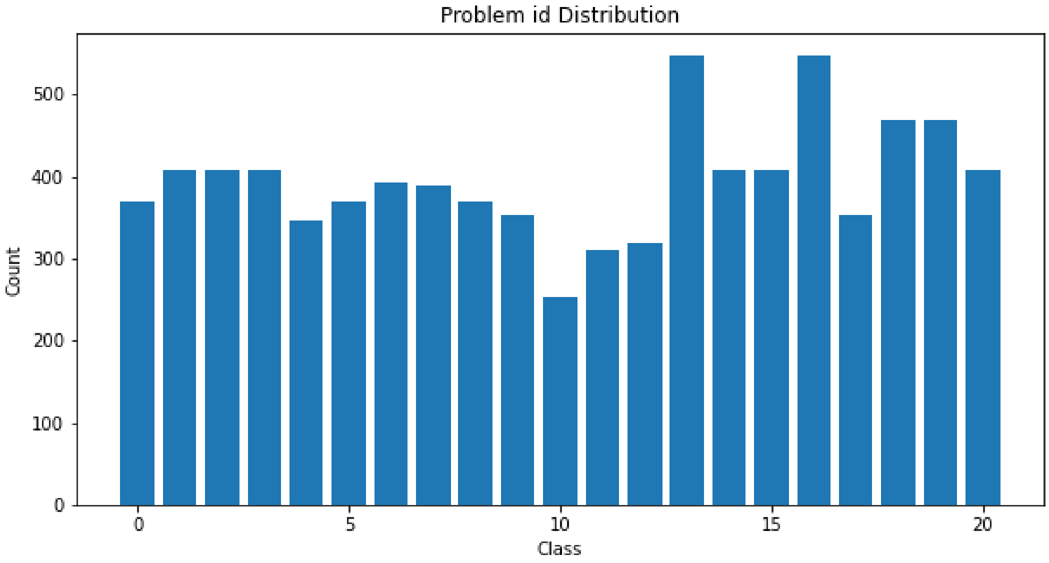

Figure 2

As we can see from Figure 2, problem id is nearly uniformly distributed, but that does not give us any information about the class balance. Thus, we proceed with checking the class distribution per problem_id

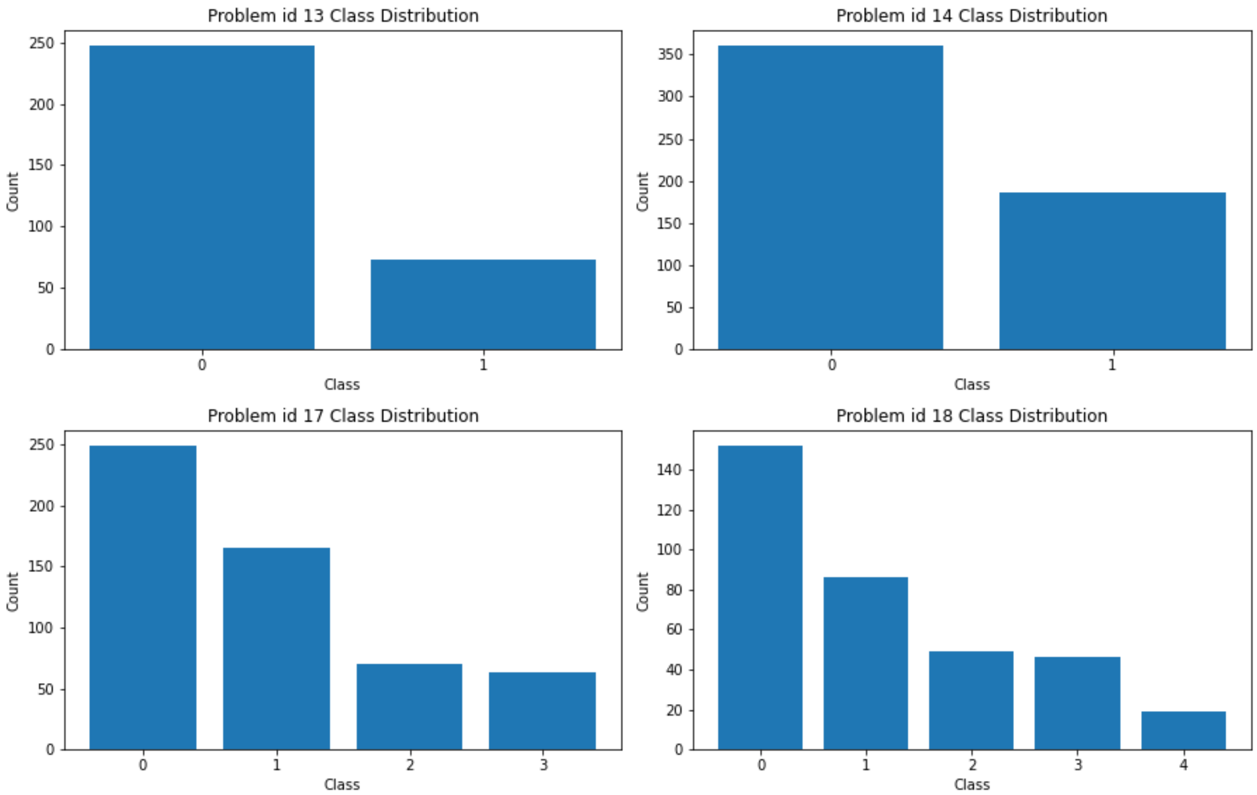

Figure 3

Figure 3 confirmed our assumption that there is also a target imbalance per problem id. This confirmation led us to our decision to use stratify splitting during cross validation for hyperparameter optimization and model performance check.

Machine Learning Methods

We decided to use tree based ensemble models XGBoost and LightGBM, as they are two of the best performing models for the tabular data that we have. Before any hyperparameter optimization we trained base models for both implementations and saw that LightGBM performed faster than XGBoost with the same performance. We discarded XGBoost from our further analyses to save resources. Next we used the Optuna framework which utilizes Bayesian Optimization for hyperparameter tuning. For tuning, we used 5-fold stratified cross validation. The stratification was applied on the column we created by combining problem_id and target to get the optimal target distribution. Finally, to predict on the testing data, we ensembled 5 different hyperparameter optimized models each using ⅘ of the data to predict class probabilities on thes testing set. We then used these 5 different probability prediction averages to come up with a final hard class prediction.

Experimental Results

As mentioned above, we used the Optuna framework for hyperparameter search on the LightGBM model. We tried to optimize the most important hyperparameters which included: max leaves, max depth, and number of estimators.

Search Range:

max_depth	Integer range 2-32
max_leaves	Integer range 30-40
n_estimators	Integer range 10-150 step_size = 10

For this search range using Optuna, the optimum hyperparameters were max_depth = 12, max_leaves = 39, n_estimators = 140. 5 fold stratified cross validation accuracy results for these values were 0.7251, and for the base model with no hyperparameter optimization, the accuracy was 0.7145. Public test set accuracy results for these models were 0.73774 and 0.72058 respectively.

List of Responsibilities

Both parties contributed equal amounts of work to the project.

Link to the Repo

https://github.com/USF-ML2/project-outliers