Machine Learning Model Overview

For the project we had to construct a dataset using multiple resources from scratch. We managed to come up with 1100 rows. All the columns of data were carefully picked, which led to a very specific feature selection. We took the percent changes of different discrete variables between the years 2000-2014 to feed into the model. Which also allowed us to skip several feature engineering steps, i.e. categorical encoding, finding missing values and outlier detection.

Our data has low cardinality, so no binning was needed. All we had to perform is normalization. Because it’s always a good practice to normalize your data, amirite?

Our project is an imbalanced data classification problem. The additional challenge was to go through the entire target column and manually label the zip codes we knew were in fact gentrified. Unfortunately, we could not cover every single zip code. Which means the evaluation metrics were going to be off by a certain degree, and that is something to keep in mind when measuring the model’s performance.

For the algorithm we’ve selected Random Forest. It proved to be superior to other techniques we’ve tried on our dataset.

We oversampled the minority class (’gentrified’) and added some additional weights to it, by calculating the optimal ‘class weight’ metric. To further fine-tune the hyperparameters, we ran GridSearch from scikit learn. We also chose precision and recall as the evaluation metrics, because they show the model’s ability to recognize and predict the relevant instances, minority class being the relevant instance.

To track the progress, we’ve used ROC curve (Receiver Operating Characteristic) that shows how capable your model of distinguishing between classes is.

After the first few adjustments we’ve managed to improve the performance by 60%.

Minority class recall also went up from 18% to 42%, with a low precision. But knowing that our dataset wasn’t entirely labeled, we are allowing some of the false negatives to be actually true negatives. Which would automatically mean a better rate.

For the majority class, we’ve managed to achieve a very decent precision of 0.92 and 0.72 for recall, f1 score being 0.81.

We believe that if we had more than just 1100 rows of data, our model would be able to run a much more accurate prediction. Our idea is to apply the model to the upcoming 2020 Census data that is nationwide, to perfect the prediction ability and issue early warnings for what’s coming for American neighborhoods.