Research Data Scientist Take-home - Memo

Dataset and initial model

Having transformed the dataset to an appropriate format (trimming non-predictive features and those with missing entries that would be difficult to insert appropriate values for), I trained an initial Logistic Regression model on an 80/20 split training set.

Quickly evaluating the accuracy of this initial model highlighted an issue with the imbalance in the data (about a 1:2.33 ratio of hired to not-hired), resulting in the model almost always predicting a failure. This was possibly a result of my selected model features not being strongly predictive enough to overcome this bias to the population distribution; I attempted to address this by making use of class weights to penalise misclassifications of the minority class, but this did not make much of an improvement to the accuracy (although the model did now make predictions for both classes slightly more evenly). For the sake of the exercise, I moved onto building an ensemble and evaluating this systematicity mitigation, but clearly this would not be appropriate for a complete project since the accuracy of the base model would be critical.

Building an ensemble

I decided to use a relatively simple approach to build the ensemble of models, that being making use of a stratified n-fold of the training set to train n models with small differences in their weights. The reasoning behind this was that their accuracies should all be similar, as per the requirement, and this could be implemented in a timely and straightforward manner. I think perhaps I chose too large an ensemble (10), resulting in a lack of variety in the resulting models since their respective training sets could not differ too greatly (all sets having 89% of the same training data); again this is something that could be optimised for in a longer term project.

Other ideas I considered for constructing an ensemble (and would explore with longer timeline) were:

- Randomly selecting a subset of features for each model (although likely not a good idea in this case since even the entire set was not sufficiently predictive), in this way building some variability between the models but keeping most of the core of the model the same. This could also make use of a fixed set of features that are always included (maybe the most predictive features) to prevent large changes in model accuracies.
- Resampling the missing entries in each feature from that feature column's empirical distribution in the dataset on a per model basis. In this way, we can consider the dataset and trained models to be random variables we are sampling from with each model. Hopefully this approach would result in a relatively narrow distribution of model accuracy in order to be appropriate to use as an ensemble, otherwise this could potentially be

tightened by resampling features only within groupings of similar individuals rather than all individuals.

Systematicity Analysis

In order to measure and evaluate the effectiveness of an ensemble of models as a systematicity metric, I decided to investigate the distribution of model predictions (and their associated error rates) across demographic groups for the ensemble and then compare these to the single model. I ended up only having time to look at this across ethnicities, although the methodology should not vary when comparing via other variables.

Firstly, I decided to look at the difference in predicted success rate and true success rate in the test set across ethnicities. Comparing these charts between the ensemble and the single model, successfully mitigating systematicity would result in a flatter distribution of these differences, spreading the error more evenly. However, visually these distributions are very similar, with no noticeable change in ‘flatness’ of the distribution, which is an indication that there may not be enough variation between models in the ensemble. In retrospect, it would have been a good idea to quantise both the difference between these charts and the relative ‘flatness’ of distribution of errors; i.e. something like the average relative absolute difference and then an entropy-like measure of the differences (could normalise the values to make a distribution).

I then decided to look at the type I (falsely predicted success) and type II (falsely predict failure) error rates across ethnicities, which should give a bit more coverage than just the difference in success rate. Again, visually these graphs were very similar between the single model and the ensemble, indicating that the ensemble is not functioning well in mitigating systematicity. Also in retrospect, quantising the differences in these charts and their relative flatness would have been a good idea; in this case I think transforming to the distributions of both types of error across ethnicities and then using KL-divergence for difference and entropy for a flatness measure would have been a good approach.

It is worth noting that for these first two metrics we are using a supervised test set, so we are assuming that this is a ground truth (i.e. that the data is accurate and representative of the population we are modeling, and also contains no bias).

I decided it would also be worthwhile to generate a synthetic dataset by independently sampling from each feature’s empirical distribution. In this way, however, we are ignoring correlations in the data between features so it could be argued that the dataset is not realistic; this could perhaps be addressed by selective grouping features to sample from etc. However, by sampling in this way, my assumption is that an unbiased model should have an even/flat distribution of success rates across ethnicities. This dataset is of course unsupervised, so I am limited to only really being able to analyse the predicted success frequencies. Comparing the single model and the ensemble, I am again looking for a flattening in the chart to indicate a more even distribution of errors (here I am assuming our ground truth is flat so no need to look at the difference in predicted and true).

Conclusions

In conclusion, there was no sufficient evidence that the ensemble implementation I used was effective at mitigating systematicity. This may not have been helped by the fact that the single model itself was a very poor predictor, so the results observed may have been noise from the training set. As well as this, with smaller error rates overall it may have been clearer in which areas systematic errors were present and how using mitigations affect these. I also think I made a mistake in the number of models used in the ensemble (at least for how I was constructing them), since the similarity in many of these charts indicated a lack of variation between these models, which is necessary for the ensemble mitigation to work as intended.