An Investigation of Posterior Probabilities in Unbalanced Binary Classification Settings

Morgan Hinhang Wu, Matthew A. Lanham

Purdue University Krannert School of Management

wu467@purdue.edu; lanhamm@purdue.edu

Abstract

We investigated the effect of population purchase prevalence adjustments on probability forecasts used to support the assortment planning decision for sparse demand products. We investigated the performance of various predictive models on various sized and various levels of imbalance. The performance was assessed using traditional statistical performance measures, as well as with probability calibration plots, which help gauge how well the models perform with regard to the actual business purchasing behavior. Both of these measures are important when determining which model performs optimally in the case of sparse demand assortments. In this study, we have found that not rebalancing consistently leads to the best overall accuracy regardless of how imbalanced the data set is. This evidence is not as conclusive with the AUC statistic, but we found many of the AUC values to essentially be no different from one another (i.e. tied) for the three rebalancing methods we researched (no rebalance/raw, down, up). Based on the findings, for all levels of class imbalance, we recommend that this data should not be rebalanced in future modeling runs.

Introduction

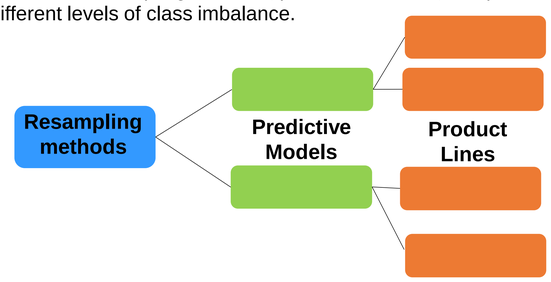
We usually use classification models as pretty reliable classifiers if we want to predict or calculate the probability that a class occurs, but we are not sure how well these classifiers perform under various level of class imbalance. When facing unbalanced data sets, traditional classifiers could create huge discrepancy between specificity and sensitivity. Luckily, there has been some research that has already realized this problem and provided some educated suggestions and directions for us to combat it - the most direct way is resampling. In this project, we would try to apply what we learned from these suggestions and perform different resampling approaches and perform evaluations on available data sets we gathered from an national retailer.

Businesses are always trying to maximize their sales or profit. But in reality there are also concerns about the different operational problems such as various costs coming from different assortment plans. And specifically to auto retailers, sparse demand patterns would also greatly effect performance of predictive models. Demands for certain product categories are much more sparse than others, and for a number of sections within a store, there could be some products without a single sell for an entire planning horizon. This specific situation makes the assortment planning very challenging and can require a different approach to picking an assortment than regression-based demand forecasts. If we treat demands as a numerical variable, the data would be expected to be very skewed with many zeros and often lead to poor predictive performance when using regression-type techniques. This situation of demands where purchasing is either a lone sell or a no-sell, leads a modeler to treat demands as a binary categorical variable (e.g. '1' for seller and '0' for non-seller).

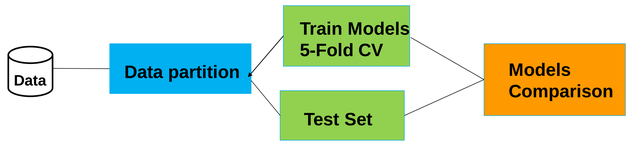
The primary issue we focus on in this research is what works best in the binary classification setting, when the datasets we use to model purchase propensity are unbalanced. This is an important concern studied by domain experts in this area. The reason for the concern is that nature of machine learning algorithms tend to favor generating probabilities toward the direction of the majority class. And to measure the effect of that, we look at how many times you classified a true seller correctly compared to a true non-seller correctly, are these statistical performance measures close or do they vary widely? We would show the results in different forms and give suggestion based on the results.

Methodology

To compare and evaluate different resampling methods, we measure the performance of combinations of resampling methods, predictive models, and product lines, which contain different levels of class imbalance.



All the predictive models were trained and tested using a 75/25 percent train/test split. After partitioning the data into training and test sets, we resampled as stated above into raw, up, or down training sets which used to train each model. As models were trained, 5-fold cross-validation was performed. This is done to reduce variability in how the model is learned, rather than using one data set. We choose 5-fold specifically to improve runtime performance compared to 10-fold.



The traditional statistical performance measures we captured were specificity, sensitivity, accuracy, AUC, Mathews Correlation Coefficient (mcc), Cohen's Kappa, and F1. We also used probability calibration plots (Kuhn and Johnson 2013) to see how the models performed with respect to the business (i.e. sell vs. no sell).

The probability calibration plot shows the predictive probabilities binned on the x-axis and the corresponding proportion of 1s (or sellers) on the y-axis for each bin. In theory, if the probabilities are calibrated well, the average proportion of 1s in each bin should follow a 45-degree line.

Since the response in our study is sold vs not sold, we consider this a business performance measure rather than the traditional statistical performance measures (e.g. AUC). Lanham and Badinelli (2015) have shown that such plots can be very useful when evaluating various models as statistical performance measures can be very similar and non-discriminatory in showing which model does provide the best decision-support.

Data & Models

The data used for this project was provided by a data science department of a collaborating retailer. Each set of similar products could be modeled together and each set had varying degrees of class imbalance (18% to 94%).

We selected five machine learning techniques that are popular for classification modeling or were implemented from other studies we examined. The methods include (1) logistic regression, (2) linear discriminate analysis (LDA), (3) quadratic discriminant analysis (QDA), (4) C5.0 decision tree, and (5) random forests.

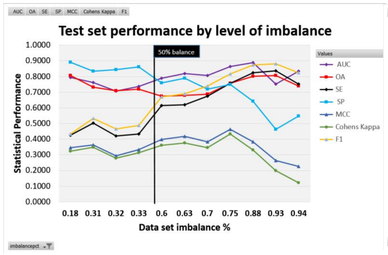
Results

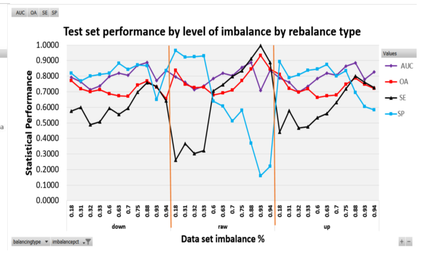
Table for Test set AUC

Test AUC	C5.0			C5.0 Total	glm			glm Total	lda			lda Total	qda			qda Total	ranger			ranger Total
	down	raw	up		down	raw	up		down	raw	up		down	raw	up		down	raw	up	
68	0.850	0.720	0.820	0.797	0.810	0.820	0.810	0.813	0.780	0.745	0.780	0.768	0.740	0.740	0.810	0.754	0.845	0.840	0.815	0.833
N	0.840	0.580	0.810	0.743	0.780	0.790	0.780	0.783	0.740	0.680	0.750	0.723	0.670	0.670		0.670	0.830	0.820	0.780	0.810
Y	0.860	0.860	0.830	0.850	0.840	0.850	0.840	0.843	0.820	0.810	0.810	0.813	0.810	0.810	0.810	0.810	0.860	0.860	0.850	0.857
89	0.810	0.810	0.820	0.813	0.820	0.820	0.820	0.820	0.790	0.780	0.790	0.787	0.790	0.790	0.790	0.790	0.820	0.820	0.810	0.817
Y	0.810	0.810	0.820	0.813	0.820	0.820	0.820	0.820	0.790	0.780	0.790	0.787	0.790	0.790	0.790	0.790	0.820	0.820	0.810	0.817
131	0.835	0.835	0.835	0.835	0.820	0.820	0.820	0.820	0.765	0.760	0.770	0.765	0.790	0.760	0.755	0.764	0.850	0.835	0.835	0.838
N	0.820	0.820	0.820	0.820	0.810	0.810	0.810	0.810	0.760	0.750	0.760	0.757		0.730	0.720	0.725		0.820	0.820	0.820
Y	0.850	0.850	0.850	0.850	0.830	0.830	0.830	0.830	0.770	0.770	0.780	0.773	0.790	0.790	0.790	0.790	0.850	0.850	0.850	0.850
174	0.890	0.880	0.880	0.883	0.895	0.895	0.895	0.895	0.855	0.840	0.855	0.850	0.870	0.855	0.870	0.863	0.885	0.890	0.880	0.885
N	0.880	0.850	0.870	0.867	0.880	0.880	0.880	0.880	0.850	0.840	0.850	0.847		0.840		0.840	0.870	0.870	0.860	0.867
Y	0.900	0.910	0.890	0.900	0.910	0.910	0.910	0.910	0.860	0.840	0.860	0.853	0.870	0.870	0.870	0.870	0.900	0.910	0.900	0.903
220	0.790	0.795	0.780	0.788	0.780	0.775	0.780	0.778	0.770	0.775	0.770	0.772	0.770	0.755	0.770	0.763	0.785	0.790	0.775	0.783
N	0.770	0.770	0.770	0.770	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760		0.740		0.740	0.760	0.770	0.750	0.760
Y	0.810	0.820	0.790	0.807	0.800	0.790	0.800	0.797	0.780	0.790	0.780	0.783	0.770	0.770	0.770	0.770	0.810	0.810	0.800	0.807
397	0.755	0.755	0.755	0.755	0.720	0.720	0.720	0.720	0.700	0.705	0.695	0.700	0.690	0.690	0.685	0.688	0.755	0.755	0.725	0.745
N	0.740	0.740	0.740	0.740	0.710	0.710	0.710	0.710	0.700	0.700	0.690	0.697	0.670	0.670	0.660	0.667	0.740	0.740	0.690	0.723
Y	0.770	0.770	0.770	0.770	0.730	0.730	0.730	0.730	0.700	0.710	0.700	0.703	0.710	0.710	0.710	0.710	0.770	0.770	0.760	0.767
Grand Total	0.825	0.798	0.815	0.812	0.806	0.807	0.806	0.807	0.775	0.766	0.775	0.772	0.760	0.763	0.758	0.763	0.821	0.825	0.806	0.816
Count of winners	8	8	6		9	10	9		7	4	7		7	7	6		8	9	1	

We highlight in red which rebalance technique led to the greatest test AUC for each model-resample method. Looking at only those models that performed the best and would actually be used, the C5.0 performed slightly better when down-sampling, the random forest permed best when not rebalancing, and the logistic regression when not rebalancing. But generally speaking, to compare between resampling methods, there is not a clear difference in AUC.

We exam different performance measures based on class imbalance levels.



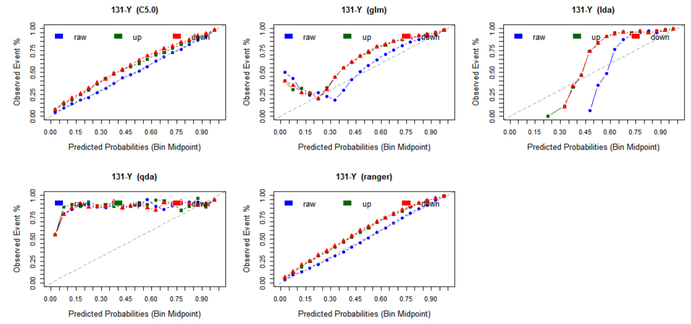


We will argue that even though sensitivity was not statistically different on average from the one-way ANOVA test, it is clear from the plots above that it is not consistency across levels of imbalance.

We perform simple linear regression on the probability calibration plots and explore the estimated slope and R-squared statistics for each of each group. While the slope provides us an idea of how well we are over- or under-forecasting certain modeling sets, the R-squared statistic tells us how close our points are to that estimated slope/line. The R-squared and slope statistics are consistent and lead to the same conclusion that not rebalancing is the preferred approach to obtain the best calibrated probabilities.

AVG of R^2	C5.0			C5.0 Total	glm			glm Total	lda			lda Total	qda	ranger			ranger Total
	down	raw	up		down	raw	up		down	raw	up			down	raw	up	
68	0.703	0.944	0.628	0.758	0.597	0.781	0.546	0.641	0.542	0.913	0.639	0.698	0.218	0.715	0.940	0.894	0.849
N	0.664	1.000	0.510	0.725	0.547	0.644	0.460	0.550	0.664	0.956	0.739	0.786					
Y	0.742	0.888	0.746	0.792	0.648	0.918	0.633	0.733	0.421	0.871	0.539	0.610	0.218	0.715	0.940	0.894	0.849
89	0.660	0.834	0.956	0.817	0.902	0.970	0.895	0.922	0.851	0.951	0.837	0.880	0.134	0.956	0.980	0.978	0.971
Y	0.660	0.834	0.956	0.817	0.902	0.970	0.895	0.922	0.851	0.951	0.837	0.880	0.134	0.956	0.980	0.978	0.971
131	0.992	0.997	0.992	0.993	0.863	0.781	0.869	0.838	0.615	0.612	0.670	0.633	0.299	0.989	0.998	0.991	0.993
N	0.994	0.995	0.990	0.993	0.834	0.772	0.836	0.814	0.542	0.469	0.586	0.532					
Y	0.990	0.999	0.993	0.994	0.893	0.790	0.902	0.862	0.689	0.755	0.754	0.733	0.299	0.989	0.998	0.991	0.993
174	0.883	0.905	0.842	0.877	0.788	0.900	0.782	0.823	0.767	0.627	0.751	0.715	0.339	0.906	0.800	0.871	0.859
N	0.862	0.921	0.868	0.884	0.760	0.913	0.749	0.807	0.677	0.580	0.632	0.630					
Y	0.903	0.889	0.817	0.869	0.815	0.887	0.816	0.839	0.857	0.674	0.871	0.800	0.339	0.906	0.800	0.871	0.859
220	0.918	0.977	0.963	0.953	0.921	0.942	0.954	0.939	0.971	0.927	0.957	0.951	0.121	0.867	0.976	0.940	0.928
N	0.938	0.983	0.968	0.963	0.978	0.970	0.979	0.975	0.975	0.944	0.980	0.966					
Y	0.899	0.971	0.958	0.943	0.865	0.915	0.928	0.903	0.967	0.909	0.933	0.937	0.121	0.867	0.976	0.940	0.928
397	0.945	0.988	0.968	0.967	0.636	0.799	0.754	0.730	0.762	0.701	0.863	0.776	0.242	0.905	0.996	0.950	0.950
N	0.942	0.989	0.964	0.965	0.301	0.646	0.537	0.494	0.566	0.491	0.771	0.609	0.398	0.846	0.994	0.920	0.920
Y	0.948	0.987	0.973	0.969	0.972	0.952	0.972	0.965	0.958	0.912	0.956	0.942	0.085	0.964	0.998	0.981	0.981
Total	0.868	0.950	0.886	0.901	0.774	0.852	0.791	0.806	0.742	0.774	0.782	0.766	0.228	0.892	0.955	0.939	0.929
Count winner	1	10	0		1	5	5		1	4	5			1	6	0	

As shown below in the probability calibration plots, each model can lead to significantly different conclusions about the performance of a model. The blue points (not rebalancing/raw) consistently show better business performance and thus provide additional insight in which model to choose instead of just using the statistical performance measures (e.g. AUC) in isolation.



Conclusions

Understanding when to rebalance or not is an important component of binary classification modeling. Some claim that rebalancing so as to achieve a 50/50 balance on the training set will allow the machine to learn without bias toward one class or another. However, fundamental machine learning theory states that a test set should always be representative of the training set. This catch-22 is what makes this area of study so interesting and challenging for researchers today.

We have found in our study of sparse demand products that not rebalancing consistently leads to the best overall accuracy regardless of how imbalanced the data set is. This evidence is not as conclusive with the AUC statistic, but we found many of the AUC values to essentially be no different from one another (i.e. tied) for the three rebalancing methods we researched (no rebalance/raw, down, up).

Using traditional statistical performance measures does not always provide the best insight into which model performs best with regard to the business. The probability calibration plots we generated for each line-application set suggest that not rebalancing leads to the "best" calibrated set of probabilities regardless of model chosen. This is because the raw sets had average slopes closest to 1 (i.e. closest to the 45% line), and average R-squared statistics closest to 1 (i.e. closer together/less variation).

References

See paper handout for additional details

Acknowledgements

We would like to thank our industry partner for this opportunity and the outstanding guidance we received from their data science team.