Credit Risk Analysis

- Gregory L. Smith
- 5/10/2022

The Credit Risk Analysis project was designed to compare different Machine Learning algorithms for their efficacy in predicting the risk of giving credit based on a field of 86 features.

In this paper we present the overview of the analysis, the results, and a summary describing the outcome.

Overview of the Analysis

In the first Notebook (credit_risk_resampling.ipynb), `LogisticRegression` was employed with several resampling models. Both Oversampling and Undersampling were studied:

1. RandomOversampling (oversampling)
2. SMOTE (oversampling)
3. ClusterCentroids (undersampling)
4. SMOTTEEN (undersmapling)

In the second Notebook (credit_risk_ensemble.ipynb), two ensemble classifiers were studied:

1. BalancedRandomForestClassifier
2. EasyEnsembleClassifier

In each of the 6 studies, metrics were computed to determine the suitability of the ML model for prediction of credit risk.

1. Accuracy
2. Confusion Matrix
3. Imbalanced Classification Report

Also, for each of the methods, the data were imported, cleaned, encoded, and scaled in preparation for each model.

Results

In each of the six models, several metrics were gathered. The following selected metrics are presented here to make comparison easier.

1. Balanced Accuracy Score - the number of correct predictions
2. Confusion Matrix - a matrix of true/false positives/negatives
3. Precision - for each of the two outcomes (high_risk / low_risk), ratio of true positives over the sum of true positives and false positives
4. Recall - for each of the two outcomes (high_risk / low_risk), ratio of true positives over the sum of the true positives and false negatives

5. F1 - for each of the two outcomes (high_risk / low_risk)the harmonic mean - a single value to show the 'goodness' of fit.

Comparison of Metrics

Model	Balanced Accuracy Score	Confusion Matrix	Precision	Recall	F1
			high/low risk	high/low risk	high/low risk

Random Over Sampler	0.6698	`[69,32]`
		`[5876,11228]`
			0.01/1.00	0.68/0.66	0.02/0.79

SMOTE Oversampling	0.6414	`[61,40]`
		`[5493,11611]`
			0.01/1.00	0.60/0.68	0.02/0.81

Cluster Centroids	0.5431	`[69,32]`
		`[10208,6896]`
			0.01/1.00	0.68/0.40	0.01/0.57

SMOTEENN	0.6435	`[77,24]`
		`[8128,8976]`
			0.01/1.00	0.76/0.52	0.02/0.69

Balanced Random Forest Classifier	0.9996	`[101,0]`
		`[11,17093]`
			0.90/1.00	1.00/1.00	0.95/1.00

Easy Ensemble Classifier	1.0	`[101,0]`
		`[0,17104]`
			1.0/1.0	1.0/1.0	1.0/1.0

Top Features (BalancedRandomForestClassifier)

The Top Features is a ranking of which features were most impactful in the predictive model. In a future study we might eliminate the less impactful features as they are noise and may pollute the model.

```
(0.32620489931446744, 'loan_status_high_risk'),
 (0.28266143696094514, 'loan_status_low_risk'),
 (0.03832357965844151, 'total_pymnt'),
 (0.037195199124358674, 'total_rec_int'),
 (0.03427425942688173, 'last_pymnt_amnt'),
 (0.031746970595770425, 'total_rec_prncp'),
 (0.02981692443571475, 'total_pymnt_inv'),
 (0.010972202088815648, 'int_rate'),
 (0.006562252554352294, 'issue_d_Jan-2019'),
```

Summary

Linear Regression - not so good

From the comparison table, above, we can see that the four Linear Regression models scored about the same.

The precision for all four sampling methods always resulted in 0.01 for "high_risk". The recall was always around 0.65 for both low and high risk. And the F1 (harmonic) was always around 0.02 for "high_risk" and 0.57 up to 0.801 for "low_risk".

Regardless of how we oversampled or undersampled the data, Linear Regression is just not a good model for this study.

It may be that the features are not "linearly separable." That is, it is not possible to draw a staight line through the data points to generate an accurate prediction.

Random Forest and Ensemble Classifiers

On the other hand, the Random Forest and Ensemble Classifiers performed quite well. The Random Forest model scored nearly perfectly in all cases. But the Ensemble Classifier matched the data perfectly.

Future Work

The results of the Balanced Random Forest Classifier "Top Features" might be used to prune the data of unnecessary features. This might improve the performance of the models and make other models a fair competitor to the Ensemble model. This will be the subject of a future study.

Recommendations

There is often a trade-off between speed and accuracy. In this case, the higher-speed models (Linear Regression with Over- and Undersampling) performed so poorly as to be unusable. In nearly all cases, the model was not able to predict a high-risk credit application.

The Random Forest and Ensemble models take more time to compute their results. But, their accuracy in detecting both high- and low-risk credit applications make them the best models. Since there's practically no difference in speed between the two models, the Ensemble model should be used for credit application risk as it matches the data perfectly.