Implementing Machine Learning Algorithms to Predict Donor Status: Preliminary Work with Data from an Institution of Higher Learning

Cecilia Coulter / Paula Baingana / Pascaline Mukakamari

Identifying potential donors allows institutions of higher learning to conduct more effective fundraising campaigns. Machine learning classification algorithms can be useful in building models to predict donor status. However, when data contains imbalanced classes, like the data we used for this project, models tend to over-index the majority class, which was non-donors in this case. These results have significant implications for institutions in that they may not pursue entities that may, in fact, become donors. In order to improve the usefulness of our model, we used a resampling technique called random undersampling (RUS) to balance the data and also the area under the receiver operating characteristic curve (AUC-ROC) metric to evaluate the performance. Our final model improved its predictive power from 67% to 76%. Institutions of higher learning can use this machine learning model to more efficiently target the pool of potential donors, saving money and time. Future research will focus on improving the predictive accuracy of our model by exploring other data manipulation techniques that minimize the effect of imbalanced data, changing thresholds for classification algorithms, and using genetic programming and feature engineering.

Implementación de algoritmos de aprendizaje automático para predecir el estado del donante

La identificación de posibles donantes permite a las instituciones de educación superior realizar campañas de recaudación de fondos más efectivas. Los algoritmos de clasificación de aprendizaje automático pueden ser útiles en la construcción de modelos para predecir el estado del donante. Sin embargo, cuando los datos contienen clases desequilibradas, como los datos que utilizamos para este proyecto, los modelos tienden a sobreindicar la clase mayoritaria, que en este caso eran los no donantes. Estos resultados tienen implicaciones significativas para las instituciones, ya que pueden no perseguir entidades que, de hecho, pueden convertirse en donantes. Para mejorar la utilidad de nuestro modelo, utilizamos una técnica de remuestreo llamada Random Under Sampling (RUS) para equilibrar los datos y utilizamos la métrica del área bajo la curva (AUC-ROC) para evaluar el rendimiento. Nuestro modelo final mejoró su poder predictivo del 67 % al 76 %. Las instituciones de educación superior pueden usar este modelo de aprendizaje automático para apuntar de manera más eficiente al grupo de donantes potenciales, ahorrando dinero y tiempo. La investigación futura se centrará en mejorar la precisión predictiva de nuestro modelo mediante la exploración de otras técnicas de manipulación de datos que minimicen el efecto de los datos desequilibrados, los umbrales cambiantes para los algoritmos de clasificación y el uso de la programación genética, así como la ingeniería de características.

Implementing Machine Learning Algorithms to Predict Donor Status:

Preliminary Work with Data from an Institution of Higher Learning

Cecilia Coulter, Paula Baingana, Pascaline Mukakamari
mccoulter@stthomas.edu, ptbaingana@stthomas.edu pascaline.mukakamari@stthomas.edu
Graduate Programs in Software, University of St Thomas, Minnesota, USA

Abstract

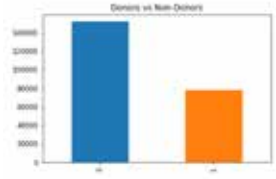
Identifying potential donors allows institutions of higher learning to conduct more effective fund-raising campaigns. Machine learning classification algorithms can be useful in building models to predict donor status. However, when data contains imbalanced classes, like the data we used for this project, models tend to over-index the majority class, which was the non-donors in this case. These results have significant implications for institutions in that they may not pursue entities that may, in fact, become donors. In order to improve the usefulness of our model, we used a resampling technique called random undersampling (RUS) to balance the data and also the area under the receiver operating characteristic curve (AUC-ROC) metric to evaluate the performance. Our final model improved its predictive power from 67% to 76%. Institutions of higher learning can use this machine learning model to more efficiently target the pool of potential donors, saving money and time. Future research will focus on improving the predictive accuracy of our model by exploring other data manipulation techniques that minimize the effect of imbalanced data, changing thresholds for classification algorithms, and using genetic programming and feature engineering.

Introduction

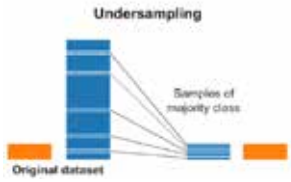
- **Problem**: What techniques can we employ with imbalanced data to develop a machine learning model that more accurately identifies potential donors?
- **Rationale**: Correcting for data imbalance before running classification algorithms should yield unbiased models that will be able to more effectively predict donor status.
- **Background**: Institutions of higher learning have plenty of data that would benefit from machine learning models that predict donor status. However, these data are sparse and imbalanced: two features that have rendered most models developed so far ineffective at predicting donor status.

Materials and Methods

- **Data**: We dealt with over 200,000 records and 76 features. The majority class was Non-donors.



- **Resampling techniques**: We used random undersampling (RUS) (pictured) and synthetic minority oversampling technique (SMOTE).



- **Classification algorithms**: We ran logistic regression, decision tree, random forest, and naive-Bayes.
- **Dimensionality reduction**: We used principal component analysis (PCA) and linear discriminant analysis (LDA).
- **Evaluation metrics**: We explored confusion matrices, recall/precision/F1 tables, and AUC-ROC (receiver operating characteristic).
- **Language**: Python.

Results

- **Preliminary results - model comparison**. Classification algorithms with raw data and 10-fold cross validation only are shown. Best model highlited in orange.

Algorithm	Accuracy with k-fold
Logistic regression	74.7%
Decision tree	74.1%
Random forest	74.4%
Naive-Bayes	72.2%

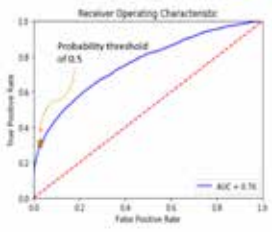
- **Logistic regression confusion matrix, no RUS**. Note the high number of false positives (in orange). With this model, most entities are predicted to be 'non-donors'.

	Predicted non-donor	Predicted donor
Actual non-donor	42927	2515
Actual donor	14997	8543

- **Logistic regression confusion matrix with RUS**. Note the improvement in false positives (in orange); fewer donors labeled as 'non-donors'. Accuracy falls to 68.77% but, at this point, we realize this metric is misleading (see next bullet).

	Predicted non-donor	Predicted donor
Actual non-donor	35463	9979
Actual donor	9349	14191

- **Evaluation - AUC-ROC for out best Model - logistic regression with RUS, our best model**; model evaluation improves to 76%.



Conclusions

- **Best model**: logistc regression, 10-fold cross validation, RUS.
- **Best Evaluation metric**: AUC-ROC.
- **Typical donor profile**: Entities with strong ties to the institution, namely, having attended the institution, being a trustee, residing in the USA and planning, giving and attending events.

Further Research

- **New algorithms**: Test new algorithms specifically designed to deal with imbalanced data.
- **New thresholds**: Change the threshold parameter to improve model's predictive power.
- **Data augmentation**: Merge with other databases, use API's and harvest more information from the existing roster.

Acknowledgements

We would like to thank Professor Michael Dorin, Clinical Professor, Graduate Programs in Software, UST, MN, for his guidance and encouragement, Dr. Manjeet Rege, Ph.D., Associate Professor, Graduate Programs in Software, UST, MN, who inspired us to undertake this project as part of a Machine Learning class and UST's Development Office for providing the data.

References

[1] Rafael Alencar. Resampling strategies for imbalanced datasets. https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets, 2017.

[2] Jason Brownlee. How and when to use roc curves and precision-recall curves for classification in python. https://machinelearningmastery.com/roc-curves-and-precision-recall..., 2018.

[3] Renuka Joshi. Accuracy, precision, recall f1 score: Interpretation of performance measures. https://blog.exsilio.com/all/accuracy-precision-recall.../, 2016.