Predicting the Downfall of Non-Profit Organizations Using Machine Learning

Author: Rubén A. Vázquez Rodríguez
Advisor: Dr. Othoniel Rodríguez Jiménez
Electrical & Computer Engineering and Computer Science

Abstract

Machine learning can be applied to finances of non-profit organizations taken from IRS Tax Forms 990ez to determine if an organization will be dissolved. Data stored on an online database was extracted, formatted, parsed and segregated using Python. The code selected the attributes that were critical, and finances were compared. Three supervised predictive algorithms, Decision Tree, K Nearest Neighbors and Naïve Bayes, were used. Results from the algorithm's predictions for organizations that were dissolved and non-dissolved are presented.

Introduction

In modern days, the amount of data available for analysis is vast enough to allow us to predict the behavior of almost anything, including image and speech recognition, medical diagnosis, traffic conditions, financial services and so on. Big data is widely being used for research and analytics. Traditional databases are not capable of handling big data. Machine learning focuses on the development of fast and efficient learning algorithms which can make predictions on data [1].

Background

Machine Learning techniques provide methods to treat and extract information from big data automatically, where human operators and experts are not able to deal with because of the level of complexity or the volume to be treated [2]. This involves finding on it the relevant information, modeling the elements, composing it and transforming it into useful information and knowledge.

Machine learning tasks are grouped into three categories: supervised, unsupervised and reinforcement learning. Supervised machine learning, used in this study, requires training with labeled data, each consisting of input value and a desired target value [1]. The supervised learning algorithm analyzes the training data and makes an inferred function. Supervised learning techniques are preferred for data analysis [3].

Problem

This study presents an application of machine learning to determine if an organization will be dissolved or not based on their finances as reported in their IRS 990ez tax forms. Knowing if the non-profit organization will be dissolved could help investors decide if it is viable to support that specific cause.

Methodology

Data Retrieval and Parsing

- Data is retrieved from Amazon Web Services S3 database and parsed from an XML format to XLSX.

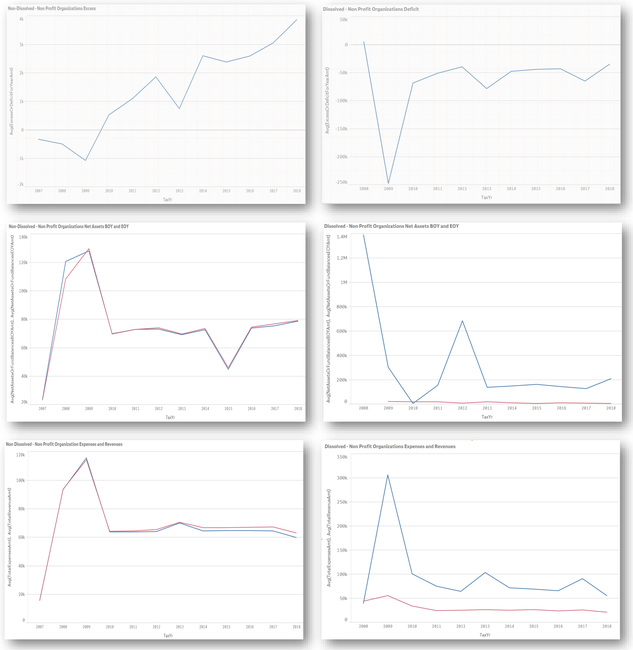
Data Segregation and Cleaning

- Two data sets were created, one for dissolved organizations and the other for the non-dissolved ones.
- The attribute “OrganizationDissolvedEtcInd” was standardized to either 0 or 1.

Attribute Selection

- Qlik tool was used to overview the data and identify significant attributes.
- Significant attributes: Total Expenses, Total Revenues, Total Deficit, Total Asset Balance at the beginning of the year and the Total Asset Balance at the end of the year

Non-Dissolved Organizations | **Dissolved Organizations**



Supervised Algorithms for Testing

- K-Nearest Neighbors, Naïve-Bayes and Decision Tree

Approach

- Train the algorithms using 80% of the subset data and 20% for testing, to determine the classifier with the best prediction accuracy.

```
# Import train_test_split function
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(features, label, test_size=0.2)
```

Code for Splitting Data into Training and Testing

Results

Once the models are trained, their accuracy can be determined using the test data. This was done by comparing the training data to the test data.

```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train.values.ravel())

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

Decision Tree Training

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Determine Model Accuracy

The code above applies to the Decision Tree. The same approach was used with the Naïve-Bayes and K-Nearest Neighbors

Model	Accuracy
K-Nearest Neighbors	99.56%
Naïve Bayes	99.28%
Decision Tree	99.61%

Model Accuracy Results

To test the algorithms two cases not used for training were chosen. First case was for a dissolved organization and the second was for a non-dissolved organization.

```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
predicted = clf.predict([[435752, 781945, -346193, 1266622, 0]])
print ("Predicted Value:", predicted)

Predicted Value: [1]
```

Decision Tree Prediction for Dissolved Organization

```
#Import Desicion Tree Classifier model
from sklearn.tree import DecisionTreeClassifier
# Import Decision Tree Classifier

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
predicted = clf.predict([[50894, 55482, -4588, 20407, 15819]])
print ("Predicted Value:", predicted)

Predicted Value: [0]
```

Decision Tree Training for Non-Dissolved Organization

Model	Dissolved	Non-Dissolved
K-Nearest Neighbors	1	1
Naïve Bayes	1	0
Decision Tree	1	0

Model Prediction Outcome

Discussion

All 3 models accurately predicted the case where the organization was dissolved. The K-Nearest Neighbors model was not able to predict the status of the dissolved organization as expected. In this algorithm, it is difficult to determine how many neighbors are required for the algorithm to accurately predict the failure of an organization. The algorithm calculated 100% precision for non-dissolved organizations and a 63% precision for the dissolved one.

Conclusions

Machine Learning can be used for predicting behaviors in different fields if there is sufficient data. The data needs to be parsed and wrangled to be able to use tools and see trends. Predicting the downfall of a Non-Profit Organization was possible using the significant attributes. There are classifiers and regression models that are more suitable for a determined type of data. In this case, the Decision Tree classifier was the most accurate at predicting if the organization was going to be dissolved or not with a 99.61%. The K-Nearest Neighbors classifier was not the best predictor since varying the number of neighbors has little to no effect on improving the outcome for the dissolved scenario.

Future Work

The scope of the study can be expanded to include all IRS tax forms including 990 and 990pf. This will need the use of a cluster to handle the amount of data available. In scenarios like this, it would be best to employ an execution framework such Apache Hadoop, Spark, Tensor Flow or Azure-ML. Also, more attributes than those selected in this work can be included in the analysis.

Acknowledgements

Special thanks to Digna Delgado, Jonathan Ortiz, Abelardo Rivera and Yeileen M. Sanchez

References

[1] S. Athmaja, M. Hanumanthappa and V. Kavitha, "A survey of machine learning algorithms for big data analytics," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-4.

[2] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", 18th International Conf. on Transparent Optical Networks, pp. 1-4, 2016.

[3] J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng, “A survey of machine learning for big data processing”, EURASIP Journal on Advances in Signal Processing, Springer, vol. 2016:67, pp. 1-16,2016.