Andrew Mevissen
October 5th, 2018

Domain Background and Problem Statement:

The domain of the project is customer support, specifically customer retention. The goal is to determine whether or not the customer will continue to use the service (churn) based on the available data. The approach that will be used is supervised learning as we will look to learn the types of customers that are likely to stay and those that are likely to leave.

Past studies have compared various supervised learning approaches such as Caruana and Niculescu-Mizil's "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics." They found that boosted trees, random forests, and unscaled neural nets were the best models [1]. Markham's "Comparing Supervised Learning Algorithms," compares the strengths and weakness of various supervised learning approaches [2]. Sabbeh's "Machine-Learning Techniques for Customer Retention: A comparative Study," was similar to Caruana and Niculescu-Mizil's in goal, but used telecom customer retention data. They also found that random forests and ADA boost had the highest accuracy at about 96%, and Multi-layer perceptrons and Support vector machines at 94% accuracy [3]. In a case study examining customer churn PayPal's "Solving Customer Churn with Machine Learning," used random forests [4].

Dataset and inputs:

The data set was downloaded from Kaggle, but originates from an IBM sample data set. There is data on 7043 customers with 21 features (including whether or not the customer was retained). The features include a customer ID, the gender of the customer, whether or not the customer is a senior citizen, whether or not the customer has a partner, whether or not the customer has dependents, the number of months the customer has used the service (tenure), whether or not the customer has phone service, if the customer has phone service does the customer have more than one line, type of internet the customer has (if any), whether or not the customer has online security, whether or not the customer has an online backup, whether or not the customer has device protection, whether or not the customer has tech support, whether or not the customer has streaming TV, whether or not the customer streams movies, the type of contract, whether or not customer uses paperless billing, the type of billing the customer uses, the amount the customer pays each month, the total amount the customer has paid, and the whether or not the customer stayed (churn).

While the data describing each customer will be used to figure out the likelihood of a customer staying, the customer ID will not be used as an input into the algorithm. We would not want the algorithm to simply learn which customers stayed based on the customer ID as we want a general solution.

Solution Statement:

Two solutions will be explored. One solution will be Scikit-Learn's Random Forest algorithm, and the other approach will be a custom neural network using Keras. These algorithms will be used predict whether or not an individual customer continued to use the company's services.

Benchmark Model and Evaluation Metrics:

The models will be compared to a guess. 73% of the people in the sample data remained and 27% churned. Guessing that the customer will remain will be correct 73% of the time; while guessing that the customer churned will be correct 27% of the time. This means that the most accurate guess is to guess that the customer will stay. Therefore, for the model to be useful it needs to produce results that are more accurate than guessing that the customer will remain.

The comparison will be done using both the accuracy and F-score of the models on a subset of testing data. The accuracy will be calculated according to Equation 1 [3], and the F-score will be calculated according to Equation 2 [5].

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad \text{Equation 1}$$
$$\text{F-Score} = (1+\beta^2)*TP/((1+\beta^2)*TP+ \beta^2*FN+FP) \quad \text{Equation 2}$$

Where TP is true positive, TN is true negative, FP is false positive and FP is false negative.

Project Design:

After the data has been loaded, it will be preprocessed into a form that can be used by both Scikit-Learn's Random Forest algorithm and the neural network. The data will be checked to make sure that no data is missing. If a sample is missing data, if it is possible to determine the value given the other known values, the determined value will be input into the sample. If it is not possible to determine the value, the sample will be dropped. The values that are categories will be one hot encoded. The data that is numerical but not categorical will be normalized. First the numerical noncategorical data will be graphed to see the full range. If the range is greater than 2 orders of magnitude, it will be normalized logarithmically. Additionally the customer ID will be removed. The churn data will be removed and one hot encoded for the labels.

After the data has been preprocessed it will then be randomly split. For the Random Forest algorithm, it will be split into a training set and a testing set. For the neural network, it will be split into a training set, a validation set, and a testing set.

For the Random Forest algorithm first a classifier will be created. Then the data will then be fit using the classifier. To optimize the model a grid search will be preformed considering the Random Forest's such as n_estimators and the max depth. The best model's accuracy and F-score will be reported.

For the neural network a model will first be created. The model will only use dense and dropout layers. The model will then be compiled and the data will be fit using the model. To optimize the model the number of hidden layers, the size of the hidden layers, activation functions, and dropout values will be explored. The best model's accuracy and F-score will be reported.

Sources:

[1] Caruana, Rich, and Alexandru Niculescu-Mizil. "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics." Proceedings of ICML'06 (n.d.): n. page. Web. http://www.niculescu-mizil.org/papers/comparison.tr.pdf

[2] Markham, Kevin. "Comparing Supervised Learning Algorithms." Data School. N.p., 25 Feb. 2015. Web. http://www.dataschool.io/comparing-supervised-learning-algorithms/.

[3] Sabbeh, Sahar. "Machine-Learning Techniques for Customer Retention: A comparative Study." International Journal of Advanced Computer Science and Applications Vol. 9, No 2, 2018: Web. http://thesai.org/Downloads/Volume9No2/Paper_38-Machine_Learning_Techniques_for_Customer_Retention.pdf

[4] "Solving Customer Churn with Machine Learning." PayPal. https://www.h2o.ai/wp-content/uploads/2017/03/Case-Studies_PayPal.pdf

[5] "F1 Score." Wikipedia. 29 September 2018. https://en.wikipedia.org/wiki/F1_score

Data:

The data was originally found on Kaggle:
https://www.kaggle.com/blastchar/telco-customer-churn

The original source of the data is IBM:
https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/