Proposal

Machine Learning Engineer Nanodegree

Capstone Proposal

Babs Khalidson, 9th October, 2022

Domain Background

Introduction

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators globally have made the industry competititive[1]. Companies are working hard to survive in this competitive market depending on multiple strategies.

There are often three main strategies for generating more revenue within a business [2]:

1. Acquiring new customers
2. Upselling existing customers
3. Increase the retention period of customers

However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy. The reason being is that retaining an existing customer costs much lower than acquiring a new one, in addition to being considered much easier than the upselling strategy. To apply the third strategy, we need to decrease the potential of customer churn by putting systems in place to do so. Hence why exploring machine learning techniques for predicting customer churn can provide huge financial benefits to companies.

Related work

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning and data mining. The majority of related work focused on applying only one method of data mining to extract knowledge, and the others focused on comparing several strategies to predict churn.

Gavril et al. [3] presented an advanced methodology of data mining to predict churn for prepaid customers using a dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively.

He et al. [4] proposed a model for prediction based on the Neural Network algorithm in order to solve the problem of customer churn in a large Chinese telecom company which contains about 5.23 million customers. The prediction accuracy standard was the overall accuracy rate, and reached 91.1%.

Idris [5] proposed an approach based on genetic programming with AdaBoost to model the churn problem in telecommunications. The model was tested on two standard data sets. One by Orange Telecom and the other by cell2cell, with 89% accuracy for the cell2cell dataset and 63% for the other one.

Huang et al. [6] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhances the process of predicting churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

Ahmad et al. [7] investigated the churn problem by developing a model using machine learning techniques on a big data platform and building a new way of feature engineering and selection. Their study measured the performance of the model using the AUC and obtained an AUC value of 93.3% after applying the XGBOOST algorithm.

Lalwani et al. [8] tacked the churn prediction problem by using a gravitational search algorithm for feature selection while several machine learning models were applied, namely, logistic regression, naive bayes, support vector machine, random forest, decision trees. The highest AUC score of 84% was achieved by both Adaboost and XGboost classifiers.

Problem Statement

Customer churn is a major problem and one of the most important concerns for large companies. Due to the impact customer churn has on the revenues of companies it is important to develop means to predict customer churn.

Customer churn is a common problem for businesses that provide subscription services. It is often difficult to determine when exactly a customer is likely to churn, due to the fact that it could be caused by several reasons.

The challenge is to predict whether a customer will churn or not using 19 input features defined in the dataset. The accuracy of the model will be the main metric for determining its success.

Datasets and Inputs

This project was inspired by the kaggle, Customer Churn Prediction 2020

The training dataset contains 4250 samples. Each sample contains 19 features and 1 boolean variable `churn` which indicates the class of the sample. The 19 input features and 1 target variable are:

File Descriptions

All of these files can be found in `dataset` folder within the submission:

- **train.csv** - the training set. Contains 4250 lines with 20 columns. 3652 samples (85.93%) belong to class churn=no and 598 samples (14.07%) belong to class churn=*yes*
- **test.csv** - the test set. Contains 750 lines with 20 columns: the index of each sample and the 19 features (missing the target variable `churn`).
- **sampleSubmission.csv** - a sample submission file in the correct format

Data Fields

1. `state`, *string*. 2-letter code of the US state of customer residence
2. `account_length`, *numerical*. Number of months the customer has been with the current telco provider
3. `area_code`, *string*=`area_code_AAA` where AAA = 3 digit area code.
4. `international_plan`, *string*,(yes/no). The customer has international plan.
5. `voice_mail_plan`, *string*, (yes/no). The customer has voice mail plan.
6. `number_vmail_messages`, *numerical*. Number of voice-mail messages.
7. `total_day_minutes`, *numerical*. Total minutes of day calls.
8. `total_day_calls`, *numerical*. Total minutes of day calls.
9. `total_day_charge`, *numerical*. Total charge of day calls.
10. `total_eve_minutes`, *numerical*. Total minutes of evening calls.
11. `total_eve_calls`, *numerical*. Total number of evening calls.
12. `total_eve_charge`, *numerical*•. Total charge of evening calls.
13. `total_night_minutes`, *numerical*. Total minutes of night calls.
14. `total_night_calls`, *numerical*. Total number of night calls.
15. `total_night_charge`, *numerical*. Total charge of night calls.
16. `total_intl_minutes`, *numerical*. Total minutes of international calls.
17. `total_intl_calls`, *numerical*. Total number of international calls.
18. `total_intl_charge`, *numerical*. Total charge of international calls
19. `number_customer_service_calls`, *numerical*. Number of calls to customer service
20. `churn`, *categorical*, (yes/no). Customer churn - target variable.

Solution Statement

With the advancements in artificial intelligence, the possibility to predict customer churn has increased significantly.

Having looked at the current literature, most research has been focused on using single estimators for making predictions. This project intends on taking an alternative approach by building an ensemble voting classifier comprised of 3 models that have the highest AUC. I will use AWS AutoGluon to find the top 3 models. The models that I will use are:

1. **LightGBM:** an implementation of the gradient-boosted trees algorithm that adds two novel techniques for improved efficiency and scalability: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).
2. **CatBoost:** an implementation of the gradient-boosted trees algorithm that introduces ordered boosting and an innovative algorithm for processing categorical features.
3. **XGBoost:** an implementation of the gradient-boosted trees algorithm that combines an ensemble of estimates from a set of simpler and weaker models.
4. **Random Forests:** Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.
5. **K Nearest Neighbours:** a non-parametric method that uses the k nearest labelled points to assign a label to a new data point for classification or a predicted target value from the average
6. **Linear Learner Algorithm** learns a linear function for regression or a linear threshold function for classification.

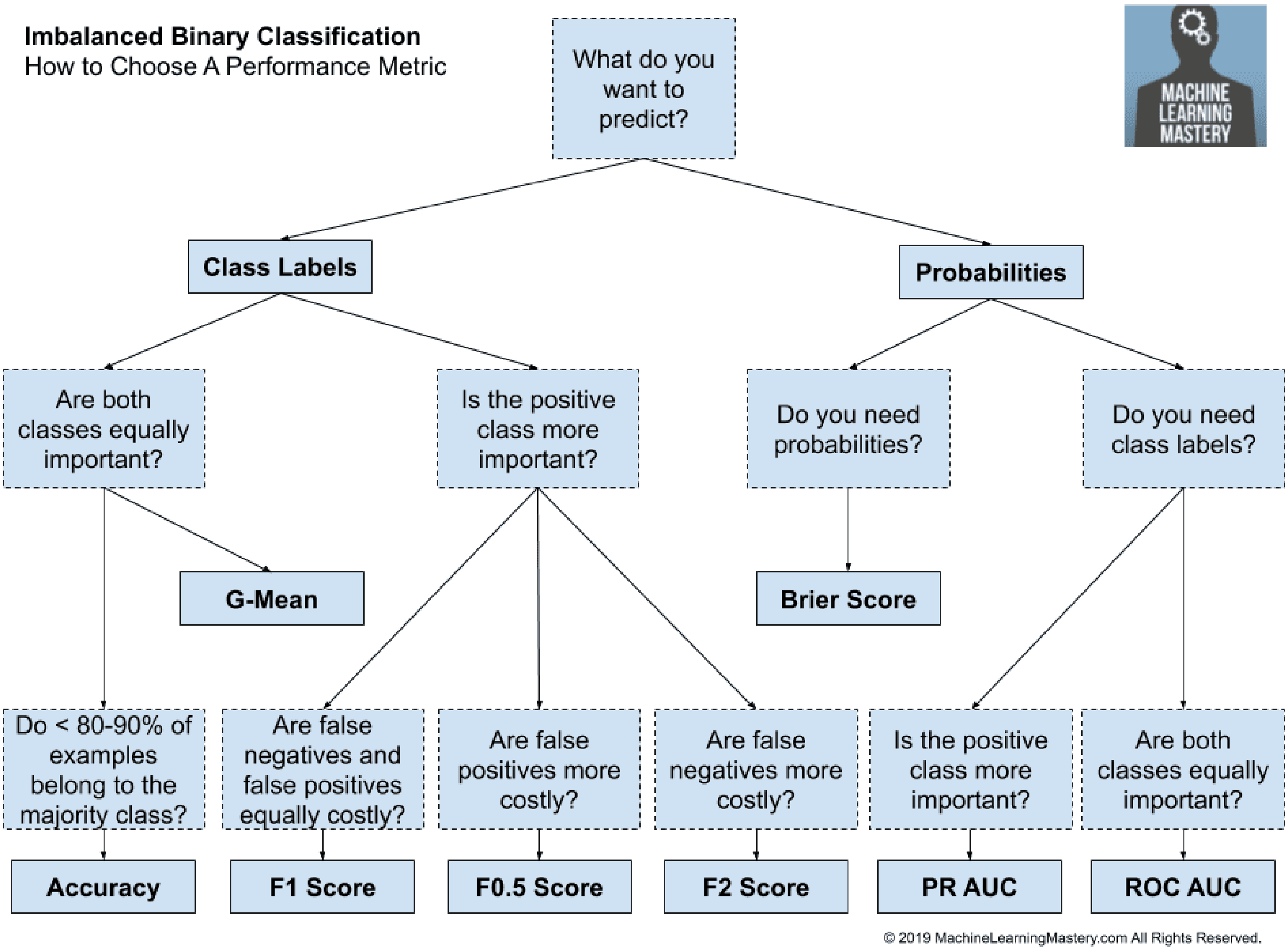
Once the best model with the highest AUC is determined the model will then be registered and deployed to an AWS SageMaker endpoint.

Benchmark Model

The baseline model that I will use to compare my model against will be a basic Linear Learner Model. This will be the minimum performance that the model will have to achieve after optimising the model. The model performance will be evaluated using the AUC score, the reason for this is explained in the next section.

Evaluation Metrics

After a quick look at the dataset, I noticed that the dataset was imbalanced. Moreover, for business reasons, the model must be able to predict each class very well, therefore the Area Under the Curve (AUC) score will be used for evaluating the model. Additionally, the decision tree diagram below further justifies why AUC score would be the most relevant metric to use to evaluate the model because we will need probabilities to determine the propensity to churn and both classes are equally important.


Imbalanced Binary Classification
How to Choose A Performance Metric
What do you want to predict?
MACHINE LEARNING MASTERY
Class Labels
Probabilities
Are both classes equally important?
Is the positive class more important?
Do you need probabilities?
Do you need class labels?
G-Mean
Brier Score
Do < 80-90% of examples belong to the majority class?
Are false negatives and false positives equally costly?
Are false positives more costly?
Are false negatives more costly?
Is the positive class more important?
Are both classes equally important?
Accuracy
F1 Score
F0.5 Score
F2 Score
PR AUC
ROC AUC
© 2019 MachineLearningMastery.com All Rights Reserved.

source

AUC score has commonly been used for churn prediction analysis as well [8][7].

Project Design

The workflow for approaching a solution:

1. `Data Analysis`: understand the datasets
2. `Features Transformation`: convert variables into features. Standardize/normalize features, apply numerical transformations
3. `Features Selection`: select relevant features
4. `Machine Learning Models`: train different models using AutoGluon. Perform hyperparameter
5. `Evaluation`: evaluate the performance of each model, and check possibilities of combining them to achieve an optimal model
6. `Deployment`: deploy the trained model to an AWS endpoint
7. `Lambda & Step Functions`: set up an AWS Lambda & Step Function for calling the deployed model.

References

1. Gerpott TJ, Rams W, Schindler A. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommun Policy. 2001;25:249–69
2. Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl. 2002;23(2):103–12.
3. Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.
4. He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.
5. Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.
6. Bingquan Huang, Mohand Tahar Kechadi, Brian Buckley, Customer churn prediction in telecommunications, Expert Systems with Applications, Volume 39, Issue 1,2012, Pages 1414-1425,
7. Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data 6, 28 (2019)
8. Lalwani, P., Mishra, M.K., Chadha, J.S. et al. Customer churn prediction system: a machine learning approach. Computing 104, 271–294 (2022)