Machine Learning Methods in Economics Course
Winter Semester 2020/21

University of Hamburg

Prof. Melanie Krause, Ph.D.

PREDICTING SURVIVAL RATES ON THE TITANIC

(Working Paper)

submitted by: Georg Zhelev

Abstract

In this project the Titanic survivor data set is used. The goal is to fit a model that accurately predicts which passengers survived by learning the relationships between the individual passenger-characteristics and their survival rate. The data set is first studied to determine what pre-processing procedures or ML methods are necessary. A simple model with the original features and a complex model with engineered features is applied to test for interactions or non-linear relationships between the variables. Data is split into training and testing using cross-validation-resampling. The models are tuned using grid search and performance is measured using accuracy score. An ensemble model combines all the models into one. A prediction with the ensemble model is submitted to Kaggle.com using the provided hold out set.

Contents

1 Introduction

The Titanic data set was chosen from Kaggle (https://www.kaggle.com/c/titanic). It is made up of the personal and travel information of the passengers of the Titanic, which sunk due to an accident in 1912. This data set is part of a prediction competition, where a separate test set is provided, which does not include target variables. For the purpose of training a prediction algorithm, the train set provided by Kaggle is used, where target values are available. By splitting this set into a smaller train and test set, the performance can be evaluated. After that the whole set is used to trian a model and predict Kaggle's test set, which does not have target values. That made submitting to Kaggle and observing the leader board possible.

Question to be Answered

The goal of the project is, based on the provided passenger information on the training set to predict which passengers survived and which did not on the hold out set. Therefore a high prediction accuracy is wanted. Understanding the underlying relationships of the data is another objective of the project. For example which types of passengers were more likely to survive the disaster and how their characteristics affect the outcome.

2 Theoretical Framework

Machine Learning (ML) methods are applied because they are particularly good for prediction. In order to assess the underlying relationship, data will be tested for interactions and non-linearity, which involves engineering features. In this case machine learning methods become handy, where statistical methods suffer from inaccurate estimation of coefficients, due to too many features per observation. Where classical statistical methods such as logistic and linear regression work with linear-only relationships, some ML methods (Tree-based Methods, Neural Networks or Support Vector Machines) can also handle highly non-linear data. Further, because machine learning provides a plethora of methods it is interesting to try as many as possible and see which one works best. Of course applying many methods on engineered features, using cross-validation for resampling and grid search to tune the models will increase computational time beyound what is a reasonable wait time. Therefore the best methods will be kept and the less performing methods discarted to keep computational time reasonable, but also to allow for a deeper theoretical undersanding of the functionality of a signle machine learning method.

Choice of Method

The chosen Titanic-data are appropriate for classification, because the variable *Survived* is binary {0,1}. That limits the choice of methods to classifiers such as k-nearest neighbors, logistic regression, decision tree, random forest, gradient boosting, a support vector machine and a feed forward neural network. Decison tree is discarted from the begining, because random forest, which is based on tree-averaging and gradient boosting, which is based on recursive-tree-bilding perform better. All models are combined in an ensemble model based on majority voting to optimize prediction performance. Methods such as linear regression, Lasso, Ridge and Elastic Net are not applied, because they are suitable for a continuous target variable.

Evaluation Criterion

The evaluation criterion is the accuracy score, which is made up of the sum of the correctly predicted true and false positives divided by all predictions. The classification type is binary (survived/did not survive). Additional model evaluation criteria are used such as sensitivity, specificity, precision and ROC curves as well as the area under the curve.

3 Methodology

Data Description

The entire Titanic passenger list is made up of the train and test samples provided by Kaggle. The train sample contains 891 observations and the test sample contains 418 observations. The total passenger number is 1309.

The data of the train set provided by Kaggle is described first. The data set contains 891 observations over 12 features. The target variable y equals 1 if the passenger survived or 0 if passenger did not survive. The features set contains two types of variables: quantitative and categorical variables. Quantitative variables are *passenger id*, *age*, *number siblings*, *number parents* and *ticket cost*. Categorical variables are *survived*, *passenger class* (1st ,2nd or 3rd), *sex*(male or female) and *boarding harbor*, *passenger name*, *cabin number*, and *ticket number*.

Kaggle provides a separate test set for the purpose of making predictions. This test set is 418 observations long and the target variable is not provided. The provided train set by Kaggle will be split into *train* and *test* sets which will be used for the training and evaluating the model. Then the model will be trained on the whole train sample and used to predict the test sample provided for the competition.

PassengerId	discrete	1-1000
Survived	binary	{0,1}
Pclass	ordinal	{1,2,3}
Name	nominal	*text*
Sex	nominal	{male,female}
Age	continuous	(0.4-80)
SibSp	discrete	
Parch	discrete	
Ticket	nominal	
Fare	continuous	
Cabin	nominal	
Embarked	nominal	

Summary Statistics

In order to determine what type of pre-processing is needed, summary statistics of the test-set-data are show below. Summary statistics are only shown for quantitative variables.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.0	891.0	891.0	714.0	891.0	891.0	891.0
mean	446.0	0.4	2.3	29.7	0.5	0.4	32.2
std	257.4	0.5	0.8	14.5	1.1	0.8	49.7
min	1.0	0.0	1.0	0.4	0.0	0.0	0.0
max	891.0	1.0	3.0	80.0	8.0	6.0	512.3

Table 1: Summary Statistics Titanic-Data

From the table it can be inferred that the PassengerId ranges from 1 to 891. Survived is between 0 & 1. Pclass (passenger class) ranges from 1st to 3rd class. Age is available for 714 out of 891 observations. The mean age is 29.7, while the minimum is 0.4(less than one 1 year old) and maximum is 80. SubSp (# siblings traveling with varies between 0 and 8. Parch (# parents traveling with) varies between 0 and 6. Fare varies between 0 and 512.3 monetary units. The distribution of the data necessitates that it should be either scaled or one hot encoded.

The mean of the survived variable is 0.38. Since survived is between 0 and 1, and the mean is more towards 0 than towards 1, more passengers did not survive. In fact looking at the sum of the 1's, from 891 observations in the train set, 342 survived which is about 40%, meaning less than half of all passengers survived. This means that the sample is slightly imbalanced toward those who did not survive (nearly 60%). This will come into consideration when splitting the sample for training and testing.

The mean age of the passengers in the train set is 29.69 years of age. From the 891 passengers there are proportionately more men (577) than women (314) in the train set. From the women 233 survived, and from the men 109. These are ratios of 74% for women and 19% of the men. This

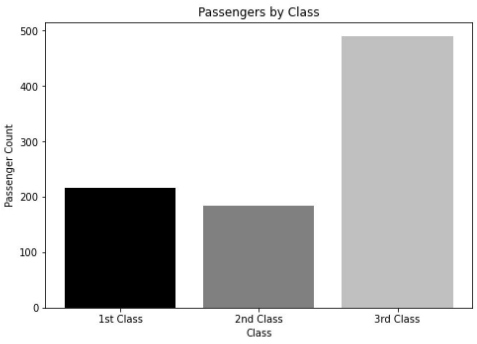

Figure 1: Passengers by Class

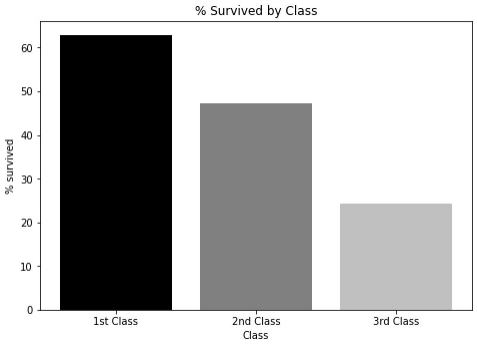

Figure 2: % Survived by Class

reflects that women and children had priority to board the limited number of life boats. When the model is trained, it will learn these relationships.

The passengers are separated into three classes. There are 216 passengers in the 1st class, 184 in the 2nd class and 491 in the 3rd class as shown in Figure 1. Figure 1 shows that passengers from higher socioeconomic classes, which booked passenger class 1 or 2 were more likely to survive than passengers in the 3rd class. The survival rate of passengers is different dependent on class. In 1st class survived 63%, in 2nd 47% and in 3rd 24% of the passengers as shown by Figure 2.

Data Pre-Processing

As inferred from the summary statistics table in Table 1, pre-processing is necessary. Pre-processing is done both on the train and the test sets provided by Kaggle. Variables that don't contribute to the survival rate such as *Passenger ID*, *Name Ticket number* and *Cabin* were dropped from the analysis. *Age* has a lot of missing values in both the train and the test set. Therefore it is imputed from the remaining values. Initially the mean age of the men and the women was used to fill in the missing values, which is too general and doesn't consider the different age groups. Therefore the *interpolate()* command from *pandas* was applied, which is a more accurate method. The variables *Pclass*, *Sex* and *Embarked* were one-hot-encoded. *Embarked* consists of the categorical acronyms for the three harbors where passengers boarded the titanic. Therefore without a numerical encoding, this information could not be used by a model. *Age*, *Fare*, *Siblings* and *Parents* have different minimums and maximums in the data provided. After the pre-processing, in the train sample remained 889 observations (down 3 from 891) and 13 variables, up 4 from original data.

	Surv	Age	Sib	Par	Fare	fem	male	C	Q	S	1st	2nd	3rd
count	889.0	889.0	889.0	889.0	889.0	889	889	889	889	889	889	889	889
mean	0.4	0.4	0.1	0.1	0.1	0.4	0.6	0.2	0.1	0.7	0.2	0.2	0.6
std	0.5	0.2	0.1	0.1	0.1	0.5	0.5	0.4	0.3	0.4	0.4	0.4	0.5
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
max	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 2: Pre-Processed Titanic-Test-Data

The test set looses no variables due to pre-processing.

Higher-Order Polynomials

Feature engineering is done to determine if non-linear combinations and interactions between variables would improve the fit. Polynomial degree of 2 is used and interactions and bias are also engineered. The feature engineering leads to an increase of the variable count from 13 to 80 variables. This is applied to both the competition test set and the train set provided by Kaggle.

Model Tuning

Parameter selection

Logistic Regression: max iterations penalty level

Lasso:

Tree:

Random Forest: depth of tree

Gradient Boosting: number of estimators learning rate

Support Vector Machine: C parameter gamma parameter

Feed Forward Neural Network: number iterations alpha regularisation parameter

Model Scoring

The data was split according to the python default which is 70% testing and 30% training samples. Train/Test split. No need for stratified split because only two classes. What are the class distributions in each split?

Making Predictions

For predictions the whole train data set is used. For the assessment of the model the train data was split into train and test. Since Kaggle proveds a hold out set with no target values, the original train split is put back together in order to use all available data for the predictions.

Predictions were made using the fitted models and appling them to the features of the testing sample. Performance was assessed based on the accuracy score.

Problem: train test accuracy 87%, validation set 78%. Model doesn't generalize good see how the classes are represented in train test, and hold out data is titanic set class-imbalanced? (may need other metric as accuracy score)

Submission after correction of expectations using cv. 81 train 82 test. A lot of coding. Not improvement in submission score in Kaggle (0.77990), only an improvement of expectations and less of a surprise. Correction of generalisation expectation is from 87 to 81, with an actual score of 78. Margin is decreased from 87-78=9 to 81-78=3. GridSearchCV carves out a validation set for the tuning of parameters from the test set.

Set is imbalanced to people who did not survived and also imbalanced to females. That is why startified cross validation was used.

4 Results

forthcoming

Evaluation of Results

Linear Regression as baseline. Forthcoming

5 Conclusion

Kaggle Submission Leaderboard

Both predictions based on the simple and on the complex model were submitted to Kaggle on 28.12.2020. The simple model scored 0.78708 on and the complex model scored 0.77033. The simple model generalizes better to unseen data than the complex model. Although the complex model hat better prediction accuracy on the test sample than the simple model. On the hold out sample by Kaggle this performance proved to be over-fitting.

The placement on the Leaderboard is 2,286 from 16,808 submissions (Top 11

Nevertheless with a score of 0.78708 there is plenty of room for improvement. Certainly when one digs into the forums of Kaggle, one can find other opportunities to improve the prediction.

Summarize question, data set method, results and how it was answered.

Positives & Negatives of Using ML Methods

Forthcoming

Appendix A: Python Code

Forthcoming

References

forthcoming