Udacity Machine Learning Nanodegree
Capstone Project

Carsten Krüger

June 15, 2018

1 Definition

1.1 Project Overview

Machine learning is used in a wide variety of fields today. Luca Talenti et al. [1] for example used a classification model to predict the severity criteria in imported malaria. In this project, machine learning will be used to build a model that can decide based on the role information of an employee whether that employee shall have access to a specific resource.

An employee that has to use a computer in order to fulfill their tasks, needs access to certain areas of software programs or access rights to execute actions such as read, write or delete a document. While working, employees may encounter that they don't have a concrete access right required to perform the task at hand. In those situations a supervisor or an administrator has to grant them access. The process of discovering that a certain access right is missing and removing that obstacle is both time-consuming and costly. A model that can predict which access rights are needed based on the current role of an employee is therefore relevant.

1.2 Problem Statement

The problem stems from the *Amazon.com Employee Access Challenge Kaggle Competition* [2] and is there described as follows:

"The objective of this competition is to build a model, learned using historical data, that will determine an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time. The model will take an employee's role information and a resource code and will return whether or not access should be granted."

This is a supervised learning problem because the dataset is labeled. Anticipated solution:

1. Explore data in order to gain insights.
2. Train many different binary classification models using standard parameters.
3. Apply transformations or regularizations.
4. Compare plain models and transformed models.
5. Pick the three best models based on the performance metric.
6. Tweak the chosen models in order to improve model performance.
7. Evaluate the tweaked models on the test set.
8. Conclusion

1.3 Metrics

To quantify model performance, the area under the ROC curve will be used. This metric is appropriate for this type of project because it works well even if the classes are not balanced. Moreover it was the metric of choice in the herein before mentioned Kaggle competition. The metric is derived by first constructing the ROC curve and then calculating the area under that curve.

"The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings" [3],

where the threshold is a value between 0 and 1 that determines how sure the model needs to be in order to classify a data entry as positive (access granted in the problem at hand). For example if the threshold was 0.7 the model would have to have calculated a probability of at least 70 % to classify a data entry as positive.

2 Analysis

2.1 Data Exploration

There are 32769 entries in the dataset with no missing values. Figure 1 shows the first five rows in the dataset.

	ACTION	RESOURCE	MGR_ID	ROLE_ROLLUP_1	ROLE_ROLLUP_2	ROLE_DEPTNAME	\
0	1	39353	85475	117961	118300	123472	
1	1	17183	1540	117961	118343	123125	
2	1	36724	14457	118219	118220	117884	
3	1	36135	5396	117961	118343	119993	
4	1	42680	5905	117929	117930	119569	

	ROLE_TITLE	ROLE_FAMILY_DESC	ROLE_FAMILY	ROLE_CODE
0	117905	117906	290919	117908
1	118536	118536	308574	118539
2	117879	267952	19721	117880
3	118321	240983	290919	118322
4	119323	123932	19793	119325

Figure 1: Top five rows in the dataset

The dataset has ten attributes. All attributes are categorial. One attribute called `RESOURCE` holds the ID of the resource for which the access has been granted or denied. There are 7518 different resources in the dataset. The target attribute is called `ACTION` and it is unbalanced because more than 94% of the rows have an `ACTION`-attribute of 1 (access granted) whereas only roughly 6% have a 0 (access denied). The accuracy metric is therefore inadequate for this dataset because even a dumb model that always predicts 1 will have a very high accuracy. The other eight columns provide role information for an employee [1]:

- `MGR_ID` - ID of the manager of employee (4243 different values)
- `ROLE_ROLLUP_1` - Role ID of employee (128 different values)
- `ROLE_ROLLUP_2` - Second role ID of employee (177 different values)
- `ROLE_DEPTNAME` - Role department description (449 different values)
- `ROLE_TITLE` - Role business title (343 different values)
- `ROLE_FAMILY` - Role family description (67 different values)
- `ROLE_FAMILY_DESC` - Extended role family description (2358 different values)
- `ROLE_CODE` - Company role code; this code is unique to each role (343 different values)

[1] `https://www.kaggle.com/c/amazon-employee-access-challenge/data`

References

[1] L1 logistic regression as a feature selection step for training stable classification trees for the prediction of severity criteria in imported malaria

Luca Talenti, Margaux Luck, Anastasia Yartseva, Nicolas Argy, Sandrine Houz, Cecilia Damon

arXiv:1511.06663 [cs.LG]

[2] Amazon.com – Employee Access Challenge

Predict an employee's access needs, given his/her job role.

https://www.kaggle.com/c/amazon-employee-access-challenge,

[3] Receiver operating characteristic

https://en.wikipedia.org/wiki/Receiver_operating_characteristic