Hillsboro Python Machine Learning Meetup

Jul/2017

Ernest Bonat, Ph.D.

Senior Software Engineer

Senior Data Scientist

DAT Wi-Fi

Username: DAT Guest

Password: beaverton dat

- 6:00 – 6:40 pm: Pizza, **water only** and networking.
- 6:40 – 6:45 pm: Welcome message by Ernest Bonat, Ph.D.
- 6:45 – 8:00 pm: Presentation and open discussions.
- 8.00 pm – 9.00 pm: Coding and learning session. Bring your Python development laptop!

Why did I create this meetup?

1. Bad traffic to Portland downtown.

2. Vert hard to find a parking lot.

3. Bad Python presentation code.

4. No time at all to review the presentation and learn something after the meeting.

We need your support:

1. Need 1 Senior Python Developers for presentation and code review every month (Co-organizers, 4-6 hours a month).

2. Email Ernest at ebonat@15itresources.com

Our Meetup Mission:

1. *"Come, Listen, Code and Learn".*

2. Finding and presenting best practices of Machine Learning using Python Data Stack.

3. Create great networking place for Hillsboro-Beaverton Data Scientists.

Meetup Schedules:

Jul/25/2017 – **Decision Trees** (non-parametric supervised learning method used for classification and regression)

Aug/29/2017 – **Random Forest** (same as Decision Trees)

Sep/26/2017 – **TensorFlow** - An open-source software library for Machine Intelligence (**Part 1**)

Oct/30/2017 – **TensorFlow** - An open-source software library for Machine Intelligence (**Part 2**)

Nov/28/2017 – **TensorFlow** - An open-source software library for Machine Intelligence (**Part 3**)

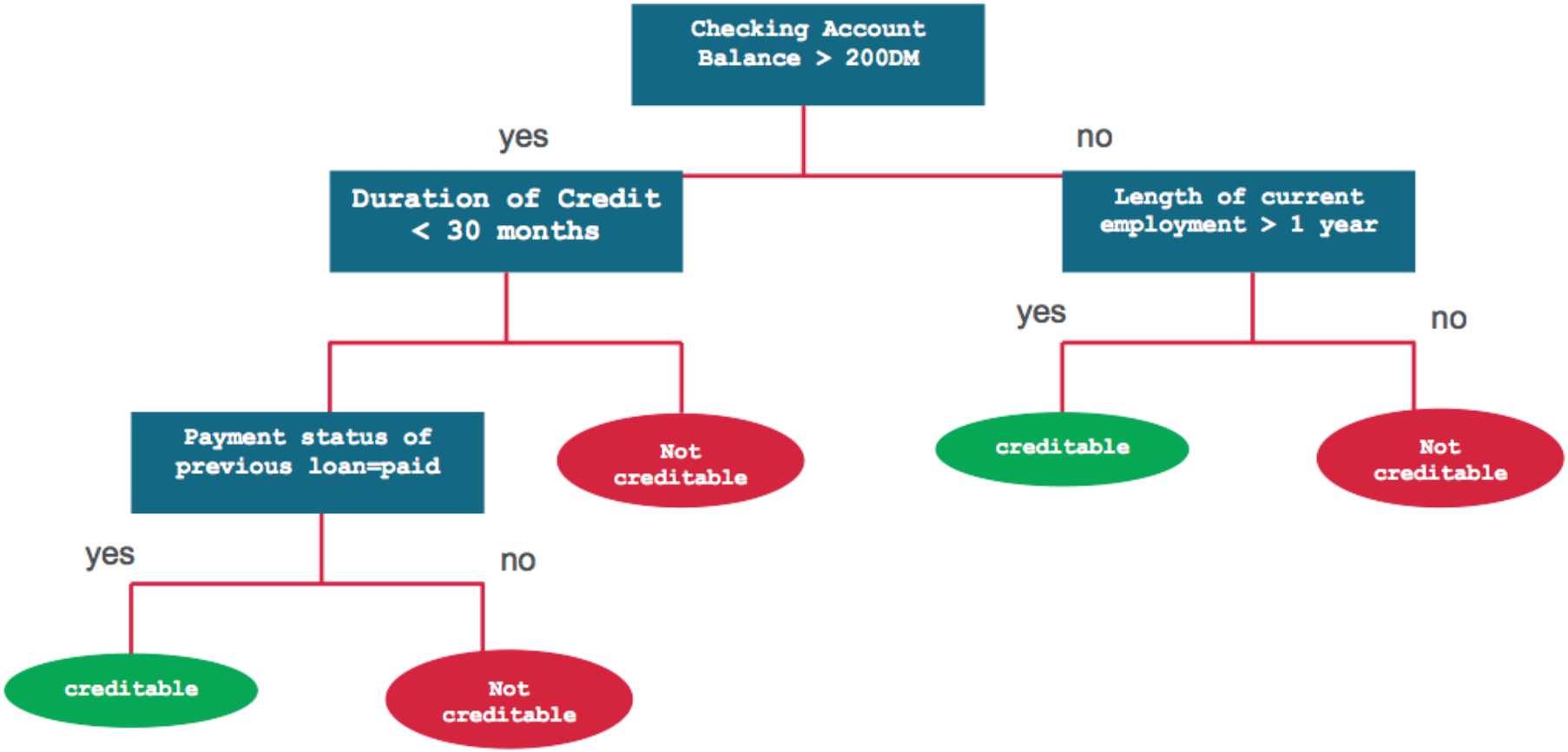
Today Presentation

"Using Decision Trees for Data Classification and Regression Analysis"

Decision Trees

Non-parametric supervised learning method used for classification and regression for categorical and continuous inputs (variables).

Example: he task is to predict the risk level of customers - *creditable* or *not creditable*.



Decision Tree algorithms that can used for classification or regression predictive modeling problems.

Classification and Regression Trees (CART Model)

In scikit-learn library we have:

sklearn.tree.DecisionTreeClassifier -> Classification

sklearn.tree.DecisionTreeRegressor -> Regression

Split the data in train and test

train_test_split(X, Y, test_size=0.2, stratify=Y)

0.2 -> 20% test and 80% training (train)

This stratify parameter makes a split so that the proportion of values in the sample produced will be the same as the proportion of values provided to parameter stratify.

For example, if variable y is a binary categorical variable with values 0 and 1 and there are 25% of zeros and 75% of ones, stratify=Y will make sure that your random split has 25% of 0's and 75% of 1's.

Decision Tree Classifier

```
DecisionTreeClassifier(criterion="gini", max_depth=5, min_samples_leaf=5)
```

Criterion – represent the quality of split:

1. "gini" criteria for Gini Index (default) - the Gini coefficient measures the inequality among values of a frequency distribution. If equal 1 expresses maximal inequality among values.

2. "entropy" for Information Gain - usually an attribute with high mutual information should be preferred to other attributes.

Maximum depth of the tree – integer number.

Min samples leaf - the minimum number of samples required to be at a leaf node.

Generate the Decision Tree graph using WebGraphviz (http://www.webgraphviz.com/)

```
from sklearn import tree
tree.export_graphviz(clf_gini, out_file='tree.dot')
```

Model Evaluation

Confusion Matrix - The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier.

Accuracy Score - this function computes subset accuracy between predicted labels (Y_pred) for a sample must exactly match the corresponding set of true labels (Y_test)

Disadvantages:

Over fitting: Over fitting is one of the most practical difficulty for decision tree models.

Look at Google for:

1. Tree Pruning
2. Bagging Ensemble Method

Training

"Business Statistics Course for Python Programmers"

(http://15itresources.com/training/)

What do we do different?

1. Corporate/in-person/hands-on training.
2. Direct business data analysis for your company needs.
3. Own and use the Ernest's Python Data Science libraries which offer full proof programming recipes.

Presentation Source Code

(https://github.com/ebonat/hillsboro_machine_learning_07_2017)