Predicting the bug fixing likelihood

Florian Spychiger

University of Zurich

December 8, 2018

Roadmap

1. Background
2. Problem Formulation & Goal
3. Data
4. Solution Approach
5. Results

Background Information

The annual cost of software bugs is estimated at $59.5 billion[1]. For the Eclipse project, there are thousands of bugs reported. An efficent bug-triaging can help developers to focus their resources and thus, save companies a lot of money.

[1] P Bhattacharya and I Neamtiu, "Fine--grained incremental learning and multi--feature tossing graphs to improve bug triaging", Software Maintenance (ICSM) 2010 (ieeexplore.ieee.org)

Problem Formulation & Goal

Problem

Bug-triaging is an important, but labor-intensive process if done manually.

Goal

Train a bug-triaging machine, which predicts whether a bug is likely to be fixed.

Raw data

The Eclipse data set can be found at
`https://github.com/ansymo/msr2013-bug_dataset`.
The raw data set consists of 12 tables:

Eclipse Bug Data Set	
reports	priority
assigned_to	product
bug_status	resolution
cc[2]	severity
component	short_desc
op_sys	version

[2]The data has been newly formated with Excel VBA.

Data model

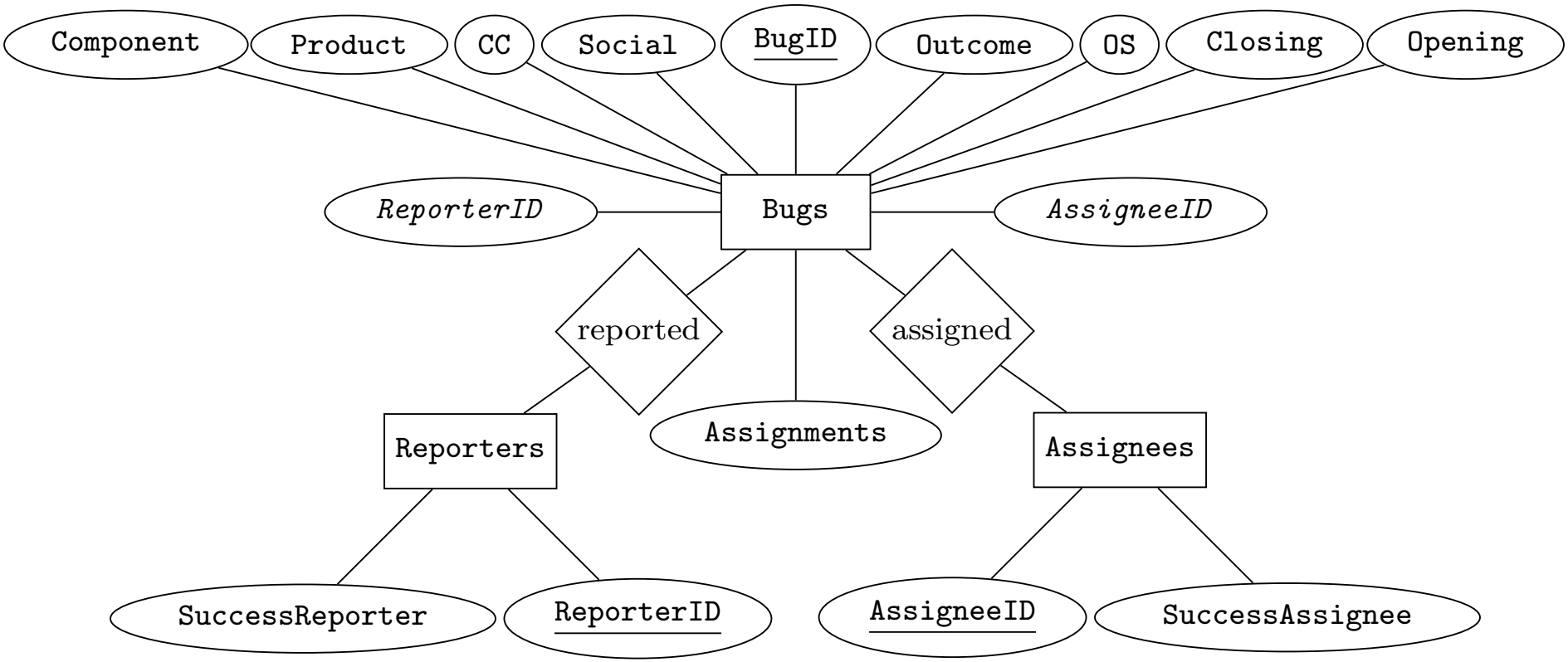

Figure: ER model of data used.

Feature Creation

Univariate Analysis

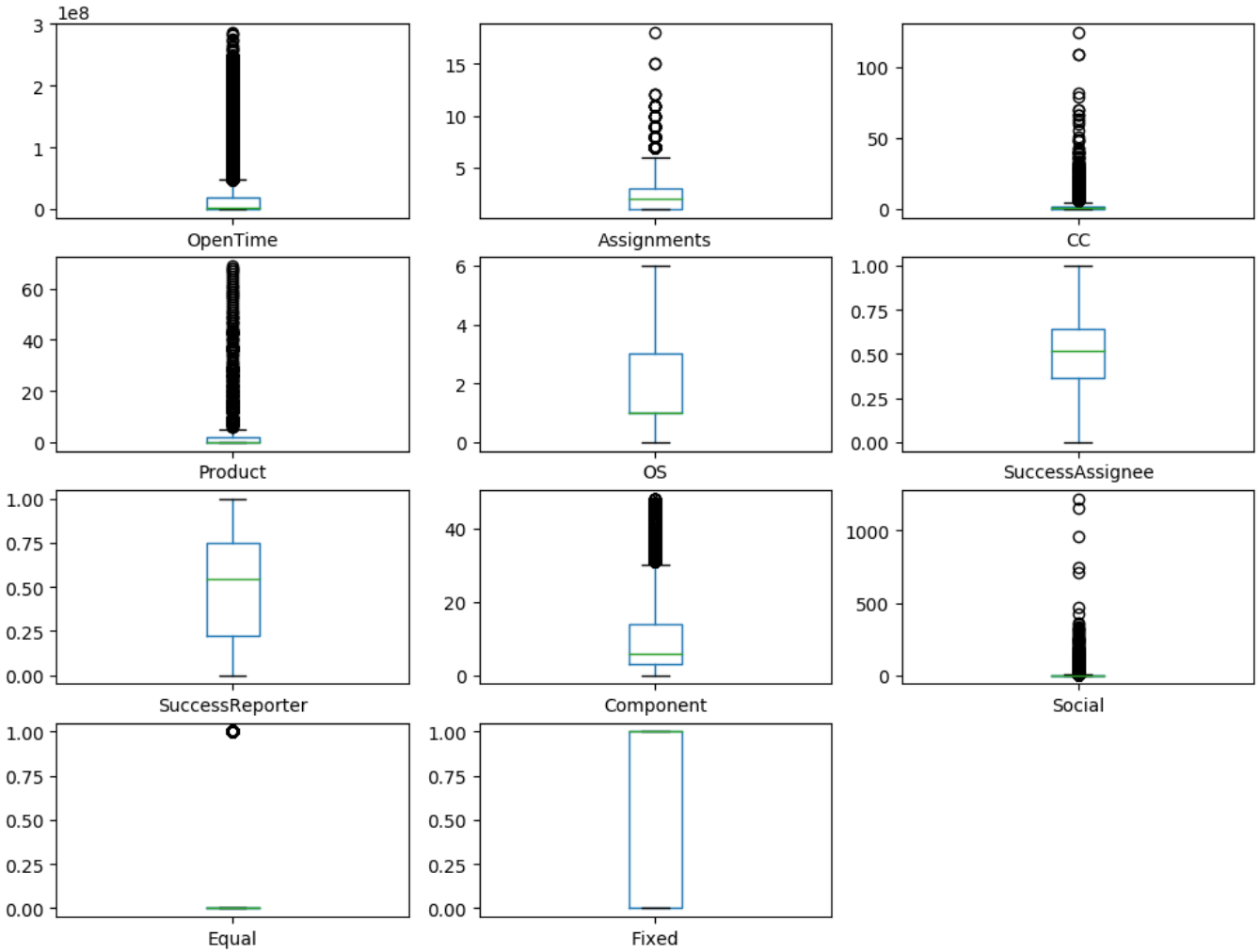

Figure: ER model of data used.

Correlation Analysis

Figure: ER model of data used.

Models

We consider 6 models:

1. Naive Bayes
2. Logistic Regression
3. Random Forest
4. Boosting Classifier
5. Support Vector Machine
6. Neural Network

We split the data set into a training (50%), a cross-validation (25%) and a test (25%) set. The training set is used to train the models and we calibrate the parametes on the cross-validation set. The final accuracy is caculated on the test set.

Accuracy

We achieve the following accuracies on the test set:

Naive Bayes	82.8098%
Logistic Regression	84.9409%
Random Forest	86.1529%
Boosting Classifier	85.4661%
Support Vector Machine	85.9105%
Neural Network	86.1125%

ROC-Curves

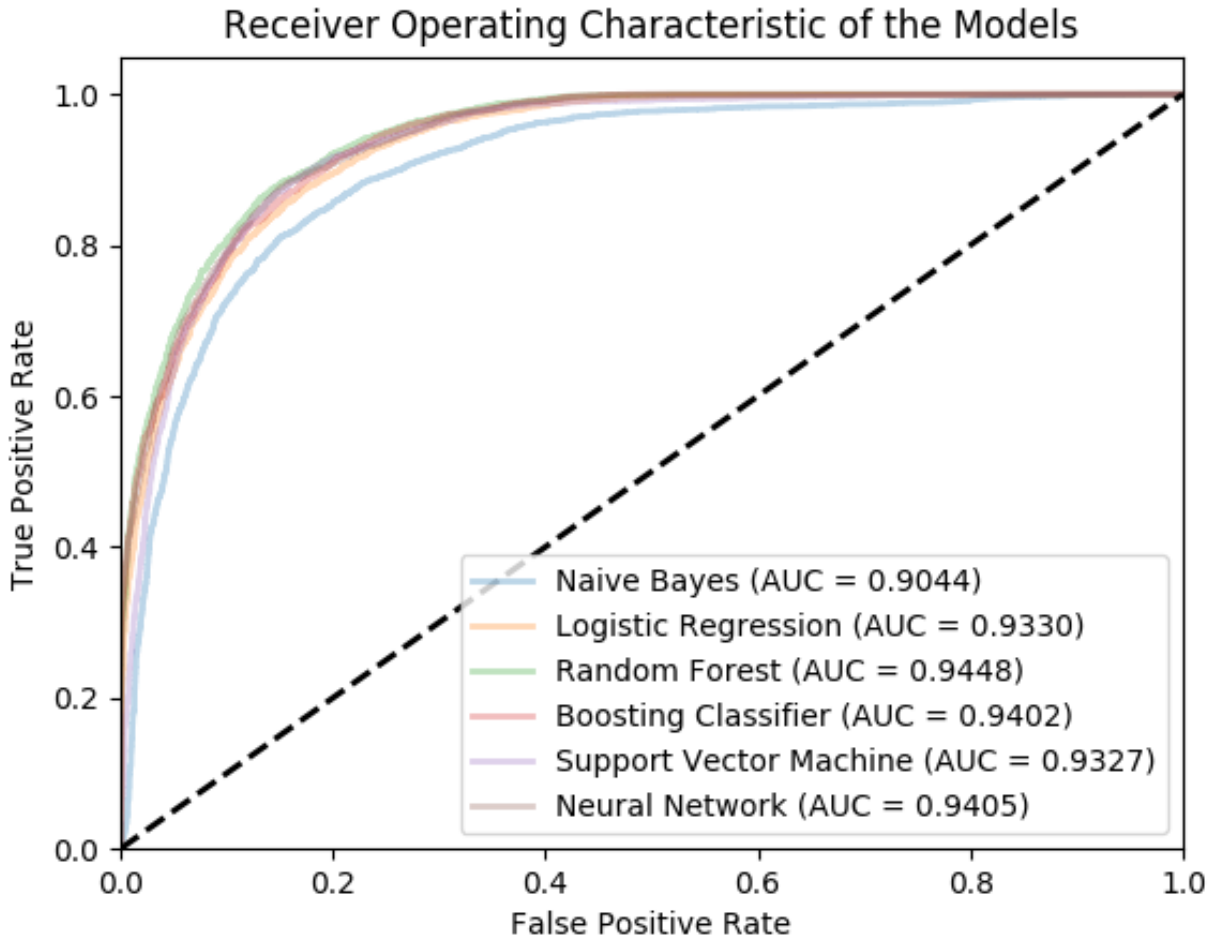

Figure: ER model of data used.