Prediction of Hard Drive Failure With SMART Stats

SUMMARY OF THE PROBLEM

Each day Sacramento Data Center uses Smartmontools to take the measurements of hard drive performance which is evaluated using SMART stats. SMART stands for *Self-Monitoring, Analysis and Reporting Technology* and is used as a monitoring standard for various attributes of the state of a given drive. Google and Microsoft performed multiple studies on disk drive characteristics in their data centers. Google found that temperature was not a good predictor of failure while Microsoft proved that there was a significant correlation.[1][2] Therefore, the goal is to determine if we can use SMART stats instead of hard drive temperature to determine whether a hard drive is going to fail before it does.

THE DATA AND THE CLIENT

The purpose of this study is to predict based on the set of SMART stats which hard drives are going to fail and compare the predictive model to the actual data collected by Backblaze data center from 2013 to 2017.[3] This will give *disk manufacturers* an opportunity to eliminate the risks that adversely affect the hard drive reliability. *Data centers* could potentially use the predictive model to replace hard drives and submit a purchase order for new equipment in a timely manner.

APPROACH

1. ***Data Wrangling and Initial Exploration:*** download the raw hard drive test data from www.blackblaze.com. The data set for each year (2013-2017) represents a separate zip file in a csv format. Use Python to import and inspect the csv files to create a cumulative data frame with all the hard drive data for the years available. Relevant columns will include the date when the measurements were taken, serial number of the drive manufacturer, hard drive model, capacity, and, most importantly, failure (boolean). The rest of the columns represent SMART stats (80 columns for 2013-2014 and 90 columns for 2015-2017). The data set would have to be cleaned due to the presence of the following characteristics:
    a. - blank fields;
    b. - inconsistent fields (SMART stats scale depends on drive manufacturers);
    c. - out-of-bound values (have to check for bounds because some drives reported to be 10+ years old);
    d. - once the hard drive failed, it was eliminated from the list, reducing the overall number of drives (count the number of drives each day);

2. ***Exploratory Data Analysis and Inferential Statistics:*** provide data visualization to find correlations that demonstrate what SMART measurements are commonly associated with failure. Form a hypothesis to test using inferential statistics.
3. ***Machine Learning:*** apply classification algorithms to predict whether the hard drive is going to fail or not based on a set of SMART characteristics. Evaluate the performance of algorithms and decide which one predicts the outcome more effectively.

DELIVERABLES

- Python code in Jupyter notebooks;

- Necessary charts and tables;
- Final report;
- PPT slides for presentation;
- Youtube video about the project (potentially)

REFERENCES

1. https://static.googleusercontent.com/media/research.google.com/en/us/archive/disk_failures.pdf
2. https://www.cs.virginia.edu/~gurumurthi/papers/acmtos13.pdf
3. https://www.backblaze.com/b2/hard-drive-test-data.html