Machine Learning Engineer Nanodegree

Capstone Project Proposal

Mark Ditsworth

December 2017

Background

The global semiconductor market is projected to be worth just under $400B in 2019[1]. The demand for semiconductors driven by digital technology has been relentless since the invention of the integrated circuit. And hot technologies such as graphical processing units, solid-state memory, and digitally controlled power electronics will continue to mandate a large volume of semiconductor production into the future.

As digital tech grows to support, regulate, and maintain our everyday lives, the importance of these devices' reliability also grows. Technologies ranging from implanted heart monitors[2] to inverters connecting utility-scale batteries to the grid[3] are all semiconductor-based technologies that can lead to dangerous outcomes in the event of failure. Thus, it is important that each semiconductor wafer that leaves the fabrication facility for sale must be functional.

However, the sheer volume of these wafers that are produced by individual fabrication prevents process engineers from running validation tests on each semiconductor die and still meeting delivery deadlines[4]. Traditional random sampling runs the risk of missing defective wafers. There is a need for a data-driven solution to detecting a failed semiconductor[5] during the course of production.

Problem Statement

Mass production of semiconductors makes individual testing of each produced IC infeasible, but the reliability of each IC must be guaranteed to prevent failure of the intended application. There must be a method for reliably classifying a device's functionality without subjecting it to individual electrical and/or thermal testing.

Dataset

[1] https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/

[2] https://spectrum.ieee.org/tech-talk/biomedical/devices/medtronic-wants-to-implant-sensors-in-everyone

[3] http://www.afr.com/business/energy/electricity/tesla-battery-responded-to-south-australian-power-failure-in-140-miliseconds-20171220-h08apx

[4] http://anysilicon.com/process-control-high-volume-semiconductor-manufacturing/

[5] B. Pavlyshenko. Machine learning, linear and Bayesian models for logistic regression in failure detection problems. 2016 IEEE International Conference on Big Data. Dec. 2016.

The dataset is obtained from the UC Irvine data repository[6]. It consists of measurements from 591 sensors scattered throughout a semiconductor chip manufacturing plant; that is, there are 591 possible features available in the data set. The 1567 measurement examples were taken during the production process of 1567 different semiconductor chips, as well as pass/fail classifications for each chip. The data is not entirely structured; there are no labels indicating what each sensor is measuring. Thus, any prior knowledge of the semiconductor fabrication process will not help in discerning which sensors would make useful features for a classifier. There are instances of missing data, which will need to be accounted for in pre-processing. Additionally, the dataset is skewed. Of the 1567 examples, only 104 are classified as failed. This will present challenges in ensuring a good balance between precision and recall of the classifier.

Solution Statement

Process engineers need a method that employs the available plant data to estimate, with a high accuracy rate, the viability of each semiconductor produced. This method will give process engineers the ability to correctly ascertain the functionality of a semiconductor wafer without individual testing or evaluation, thus allowing for high product throughput.

Benchmark Model

The benchmark model will be a binary classifier that uses a number of sensors' data as features, and predicts the state of the semiconductor wafer: functional or non-functional. The classifier will first be attempted with logistic regression. If sufficient results cannot be attained with logistic regression, more advanced classification models will be attempted: SVM, decision tree, or ensemble learners.

Evaluation Metrics

An effective classifier for predicting the functionality of a semiconductor device should have an accuracy above 90%. Accuracy is defined in terms of confusion matrix components by the equation below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

[6] M. McCann, A. Johnston. http://archive.ics.uci.edu/ml/datasets/SECOM.

Furthermore, to ensure proper functionality despite the skewedness of the data set, the Matthews Correlation Coefficient should be above 85%. The Matthews Correlation Coefficient is a measure of the quality of binary classifications, which works well on differently sized classes[7]. It is defined in terms of confusion matrix components by the equation below.

$$MCC = \frac{(TP \times TN) + (FP + FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(FN + FN)}} \times 100\%$$

It should be noted that in the event the equation above results in an undefined value, the Matthews Correlation Coefficient is defined as 0.

Project Design

To begin, the raw dataset must be processed for missing values. While one option would be removal of any example with missing sensor data, the low number of failure data necessitates the inclusion of as much of this failure data as possible. Removal of any failure data will only increase the problem of skewed data. Therefore, it may be more meaningful to replace missing data with averages. If this method for handling missing data results in a poor classification model, basic regressors will be trained to attempt to estimate the missing values.

Second, statistical analysis will be performed on the dataset to feature scale and reveal any sensors that are clearly not of use in predicting classification (e.g. the value has no variance). After the removal of any obvious non-useful sensors, principal component analysis (PCA) will be performed to further narrow down the sensors that are useful as features for the binary classifier.

Once the useful features are identified, a binary classifier will be trained with k-fold cross validation. Several different classifier models will be created and refined on the cross validation set, starting with the simplest implementation: logistic regression. If the complexity of the problem cannot be captured with logistic regression, more advanced classification models: SVM, decision tree, and ensembles, will be tried. The model that performs best on the

[7] http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

test set after optimization of parameters through grid-search will be selected as the chosen model for implementation, so long as the performance meets the evaluation criteria.