# Correlating Sensor Data with Sensor Diagnostics

Muhammad Kasim Patel, Sarthak Jagetia, Namir Fawaz

BOSTON UNIVERSITY

## Introduction

In all the IoT hype one thing that is agreed upon by most pundits is that data is the value of the IoT. Hardware devices such as sensors digitize the physical world and generate all this valuable data.

The data can be useful in lots of ways and can help determine the state of the system and how it can be improved. Thus, the objective of our project is to work with SICK in order to analyze their sensor data, find any correlation between operational data and sensor diagnostics data, and return meaningful analysis of the data.

## Problem Formulation

The problem was considered to be an open ended problem by taking the data and visualizing different parameters of the sensors and object. The goal was to find

- Correlations between different parameters of the sensor and the object data
- Look for anomalies in the data if any existed.

A few examples were given to us in the project proposal. We decided it would be interesting to search for a correlation between the speed of the conveyor belt and the dimensions of the packages.

Apart from these, there were other anomalies and relations which we found while analyzing the data and correlating different features of the dataset.

## Data Pre-Preprocessing

- For Data preprocessing, we normalized the data using the sklearn normalizer feature
- Also, every XML and TSV file had timestamps in string format. So we had to convert that to datetime object and strip the time feature so that we can merge the data frames for data analysis
- We then dropped some columns of data which had standard deviation = 0, to save space and decrease time complexity
- Furthermore data like sensor-names, device id etc were not that useful in processing and analysis so we dropped them too.

## Linear Regression

- We developed 5 different models of Linear Regression with which to correlate the data. We decided it would be interesting to search for a correlation between the speed of the conveyor belt and the dimensions of the packages. We used the csv files to load data into Pandas Dataframes, and split all data into a 70:30 ratio in order to train and test the models.
- The first model attempted to find a linear correlation between the speed of the conveyor belt and the three dimensions of the package. The second model attempted to find a linear correlation between the speed of the conveyor belt and the volume of the packages. Finally the last three models attempted to find a linear correlation between the speed of the conveyor belt and any one of the package's dimensions.

## Time-Series Analysis

- For Time-Series analysis we analyzed the different parts of the data and tried to find any correlation between them. One of them which we found worth highlighting was the relation between the speed and the volume of the boxes which were passing through the conveyor belt.
- In the plots below we can see a clear relation between the volume of the boxes and the speed of the belt.

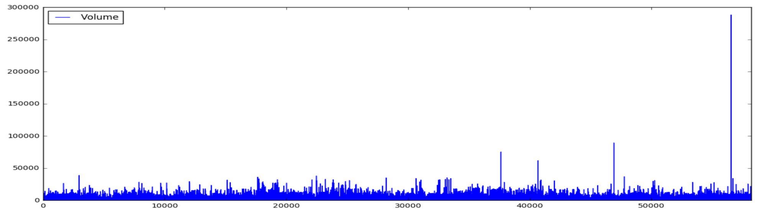



**Figure 1.** Volume vs Time

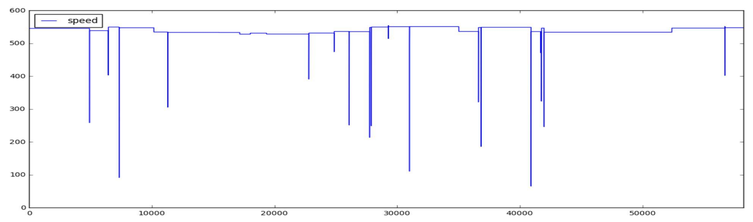
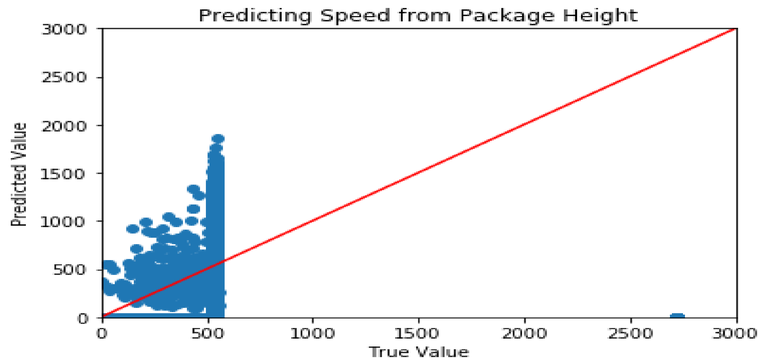



**Figure 2.** Speed vs Time




## Results

- In the linear regression, we created 5 different models; we found the model of speed vs height to have the least R-squared of 0.414
- For Time series analysis as you can see from the plots we were able to find some kind of correlation between the speed of the conveyor belt and the volume of the packages.
- Intuitively, if suddenly a number of larger volume packages are put on conveyor belt, then we should see a slight change in the speed of the conveyor belt.
- And in the KNN algorithm we found out that the algorithm did not do a good job at predicting the neighborhood values since the accuracy of the test data was less than 0.1

## Conclusions

In the end, we were able to implement many of the tools we learned over the course of the semester in a real life project. Still, we were quite unsuccessful in finding any interesting trends in the data.

We attempted to use Linear Regression, K-Nearest Neighbors, and Time Series Analysis and found distinct relations between data points such as package dimensions and conveyor belt speed using only time series analysis. In the case of our project, definitively finding no trends was equally as important as finding trends.

By finding some trends between the data points, we are able to say with confidence that given our set of sample data, the dimensions of packages (volume) should in some ways affect the speed of the conveyor belt, but not the temperature.

## References

1. A guide to appropriate use of Correlation coefficient in medical research