CAS CS506: Project Proposal # 2
Team SICK

Team Members: Namir Fawaz, Sarthak Jagetia, Muhammad Kasim Patel
Mentor: Deepak Kolippakkam

Introduction:

- The objective of our project is to work with SICK in order to analyze their sensor data and return meaningful results.
- This may sound rather open ended, and that is because there is not one specific question that SICK is looking to solve.
- Rather, they have provided us with a large two separate types of data, Operational Data (XML) and Sensor Diagnostics Data (TSV) that we will use to find noteworthy correlations.
- Instead, our main focus will be looking through data in order to create hypotheses or questions relating to different data points we believe may yield interesting results.
- As of now, we have come up with a couple questions we believe may yield very meaningful results:
    1. If voltage is spiking or or lowering, how is this affecting the operational data?
    2. If the conveyor belt is speeding up, how well is the barcode information being captured?
    3. At what speed does the sensor fails to read the barcode information?
    4. What other factors affect the operational data? Eg: temperature, lighting conditions etc.

- The first question is more open ended, because it will allow us to examine the voltage of a given sensor using the Sensor Diagnostics Data, and infer if lower or higher voltages affect any aspect of the Operational Data.
- The second question is much more specific, but is a good starting point on which to build other hypotheses or questions.
- The third and fourth questions can be useful to improve the conveyor belt efficiency
- We believe these are good starting points in order to begin work on our project. We will add few more questions as we explore the data and the project itself further.

Analysis Methods:

- As mentioned above the data provided to us is Operational Data and Sensor Diagnostics Data. They are in the formats of XML and TSV, so we can choose to not focus on collecting data.
- The aim of the project is to find correlations between attributes and to find answers to many such questions proposed above.
- In order to do that we plan to use various correlation and regression analysis methods like Linear Regression

- For that we first have to determine dependent and independent variables so that we can find how the value of the dependent variable changes when any one of the value of independent variable changes.
- The validation process can involve analyzing the goodness of fit of the regression, analyzing whether the regression residuals are random, and checking whether the model's predictive performance deteriorates substantially when applied to data that were not used in model estimation.
- We can also try and use clustering like kmeans++, GMM or Hierarchical to see if there is any trend in the data.
- This trend can then be used to formulate more questions and we can use regression models on those variables to find correlation if it exists!

Schedule/Timeline:

We are still deliberating on technical implementation. We will try out different methods and see which one of them works best for the given dataset.

A brief and tentative timeline of the milestones:

- Develop a high-level understanding of the problem and brainstorm the initial approaches by 15th November
- Investigate and implement best approach (decided by consensus and discussion with the mentor) by 28th November
- Run tests and improve the analysis by 3rd December
- Submit initial results to the mentor by 8th December
- Demo final solution by 10th December