SMU
SINGAPORE MANAGEMENT UNIVERSITY

SMU-X

SMU-X ACCT414: ACCOUNTING ANALYTICS CAPSTONE

PROJECT STATEMENT

Star Asia Trading (SAT) is an urban shoe manufacturer and a wholesale distributor of women's, men's and kid's footwear, producing shoes for major shoe brands like Adidas, Nike and Zara. SAT seeks to identify an adaptive, "smart" model where an optimal combination of known variables observed on the shop floor at any given time, can be utilized to gauge operational costs incurred.

As a result, the company strives to obtain a more complete picture of their operations at any time, by improving data utilization of data being collected on the shop floor. They have tasked our group with the developing of this **model through an open source platform** as well as a **visualisation tool which facilitates interaction with the model.**

In the context of this study, the focus of the project will be to estimate overhead costs per pair of shoes manufactured and determine the specific variables that can be used to make this estimation as accurately as possible.

PROJECT TIMELINE

WEEKS

- 2-5: DATA EXPLORATION AND CLEANING
- 3-5: ANALYSING AND PREPARING DATA FOR MODEL BUILDING
- 5-7: BUILDING AND TESTING OF MACHINE LEARNING (ML) MODELS
- 7-12: DEVELOPMENT OF VISUALISATIONS
- 10-13: FINAL REFINEMENTS TO ML MODELS AND VISUALISATIONS

BUILDING THE MODEL

CONSIDERATIONS TO THE MODEL

Data for shop floor parameters was recorded in a daily format sorted by the respective factory plants, but dependent variable (overhead cost) was recorded according to months and was representative of the factory as a whole. Our group decided to conduct analysis on two fronts.

- **Dataset by Day:** Monthly Overhead Costs would be pro-rated across each day to each plant based on daily individual output by that plant.
- **Dataset by Month:** Daily parameters by each individual plant would be aggregated together by months.

The dataset was split respectively into **training** and **testing** sets. Next, three algorithms – **linear regression, LASSO & Random Forest** were used to train the model on the training dataset. Subsequently, the trained model would then be tested for its predictive accuracy based on the testing dataset. To compare all the models, we use mean absolute error (MAE).

ALGORITHMS EXPLORED

LINEAR REGRESSION

Provides an equation that shows economic relationship between independent variables and the dependent variable. Regression models are simple to understand and thus easily communicated to management.

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

This model is suitable when we have a large number of variables, but small number of observations. It provides a mode stable equation in the long run (but with less accuracy in the short run). LASSO will shrink coefficients and select useful variables.

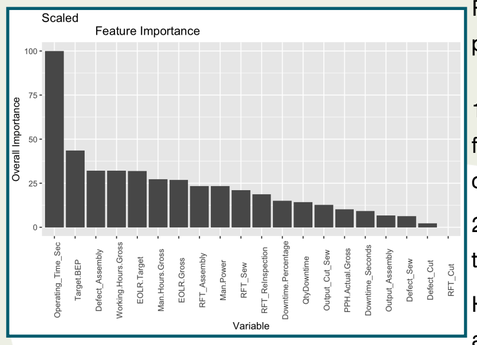
RANDOM FOREST

This model utilises an ensemble learning method, which constructs a multitude of decision trees and combines them to get more stable and accurate predictions. Random forest allows us to rank the importance of significant variables relative to one another, but it is difficult to interpret.

RESULTS OF OUR MODELS

Based on the three algorithms, **Random Forest** yielded the best MAE thus, bringing the most predictive value. We share the results of our various Random Forest models below.

HOW DOES RANDOM FOREST SELECT VARIABLES SIGNIFICANT TO THE PREDICTION?



For numerical predictions using Random Forest, variables are deemed significant based on its predictive value. It primarily does this in two ways statistically.

1) **Percentage increase in mean square error (MSE) to accuracy and the variable is permuted (IncMSE).** For each feature, the algorithm will randomly permute values of the feature and subsequently measure the resulting increase in error. The higher the increase in error, the more the predictive power of that variable.

2) **Gini Impurity.** Gini impurity is a metric used in decision trees to determine which variable to split and at what threshold to split the data into smaller groups.

Here, key variables that are used to predict overheads are **Operating Time, number of Defects, Working Hours** and the **End-of-Line Rate (EOLR).**

COMPARISON OF ACTUAL AND PREDICTED VALUES

1) TOTAL OH, DAILY FIGURES

MAE: 24.28944

2) TOTAL OH, MONTHLY FIGURES

MAE: 27030.47

3) VARIABLE OH, MONTHLY FIGURES

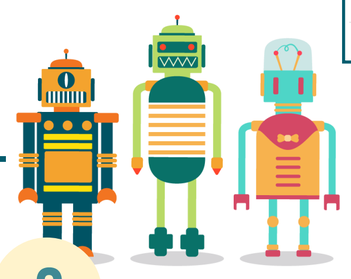
MAE: 4448.603

4) VARIABLE OH, MONTHLY FIGURES (PREDICTORS ARE NON-AGGREGATED)

MAE: 7238.86

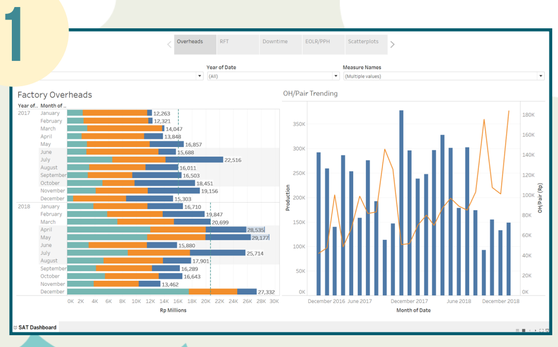
1

2 3 4



VISUALISATIONS

1



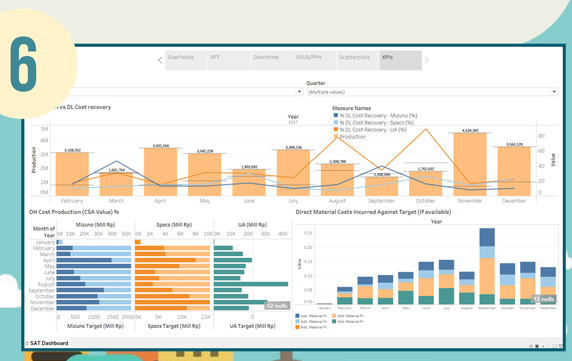
2

3

4

5

6



1) OVERHEADS

Provides a historical trend analysis of factory overheads data, categorised into Fixed Overheads, Variable Overheads and Operating Expenditures

2) RIGHT-FIRST-TIME (RFT)

RFT is an important metric used by SAT's management to measure the number of pairs that are produced without defects on the first attempt. This dashboard provides an overview of defects and RFT within the factory

3) DOWNTIME & OVERTIME

Provides a historical trend analysis of downtime and overtime in SAT's factory, with visual aids—like a heatmap, to accentuate the business units that require management's attention

4) END-OF-LINE RATE & PAIRS PER HOUR

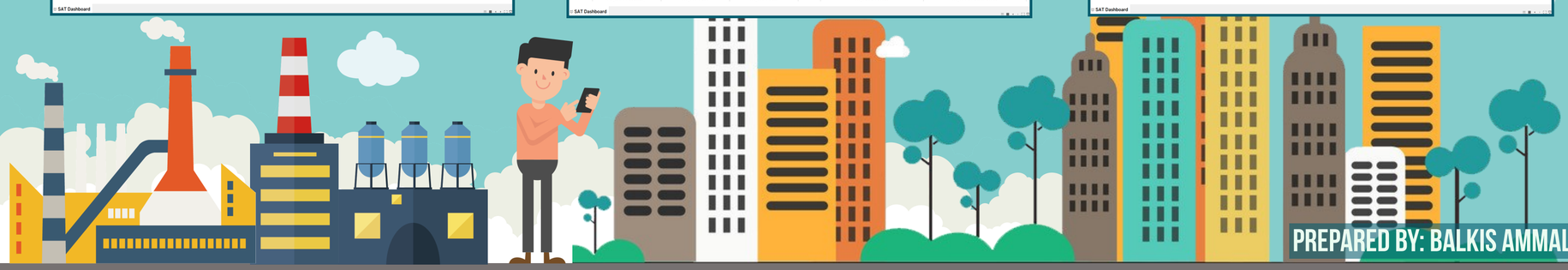
EOLR and PPH help management to compare the efficiency of each Plant and Cell within the factory. This dashboard depicts how these measures have changed over time

5) SCATTERPLOTS

Scatterplots to depict the linear correlations between certain shop-floor parameters and Overheads cost per Pair

6) KEY PERFORMANCE INDICATORS (KPIS)

Lastly, this dashboard provides a visual analysis of KPIs that are used to measure the factory's performance



PREPARED BY: BALKIS AMMAL, GAVIN ONG ZE KAI JEROME JEEVAN NAIDU, TAN YAN HAN KENNETH