BUS 4045:
CAPSTONE DATA PROJECT.
FLIP ELECTRONIC

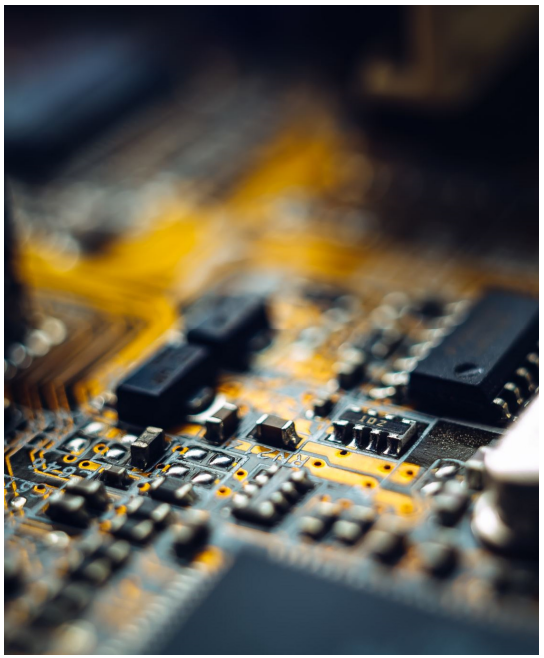
George Brown College, Toronto, Canada.

Summer 2020, AY 2019-2020.

Presentation Outline

- Background and statement of business problem.
- Data Audit Result
- Predictive Model Methodology
  - ✓Classification Algorithms
  - ✓Ensemble Learning
  - ✓Class Imbalance
- Analytical Results
  - ✓Correlation analysis
  - ✓EDA
  - ✓Gain chart/ AUC graph
- Recommendation
- Future Work

Company's Introduction



The company has a unique stocking program that provides its customers with select electronic components that fulfills the most challenging needs of the supply chain.

Middle Inventory model: Focuses on securing parts that are in short supply and high demand as well as products that are End of Life (EOL) or obsolete.

The Engagement Model

FLIP performs detailed Sale-ability Analysis

- Assess demand from existing & target customers
- Review posted demand from Open Market Distributors
- Promote through Inside Sales, Rep Network and Web Posting
- Assess current market price & recovery time

Business Challenge (Sale-ability Analysis)

A user manually visit electronic component ecommerce website to determine how many times a component has been searched. Based on this the component is classified as:

1. Cold – if it has been searched <10 times.
2. Warm – if it has been searched [10, 40] times.
3. Hot – if it has been searched >40 times.

This classification result is then used to determine whether to procure the product.

Challenge – A very time-consuming process. Automate this process.

Data Audit – Files overview

- Inventory/ Stock overview - Location of where parts are.
- Sales Invoice (Lack information about characteristics of customers) – Invoice. Contains rudimentary information such as customer name, quantity sold, price, …
- Purchase order from supplier

✓ Heat Index 1 (with 73 records). The principal file on which we worked.

- Heat Index 2 (with 1373 records – class imbalance). These were similar to Heat Index 1, but with larger number of records.

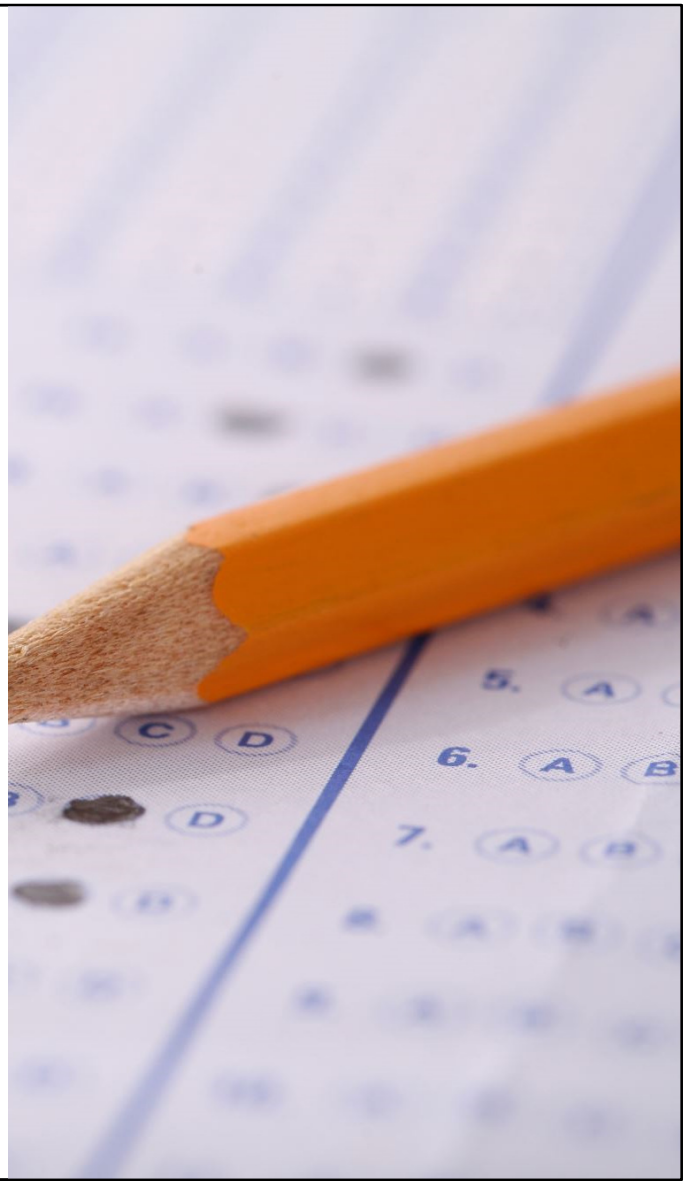
Data Audit – Missing Values

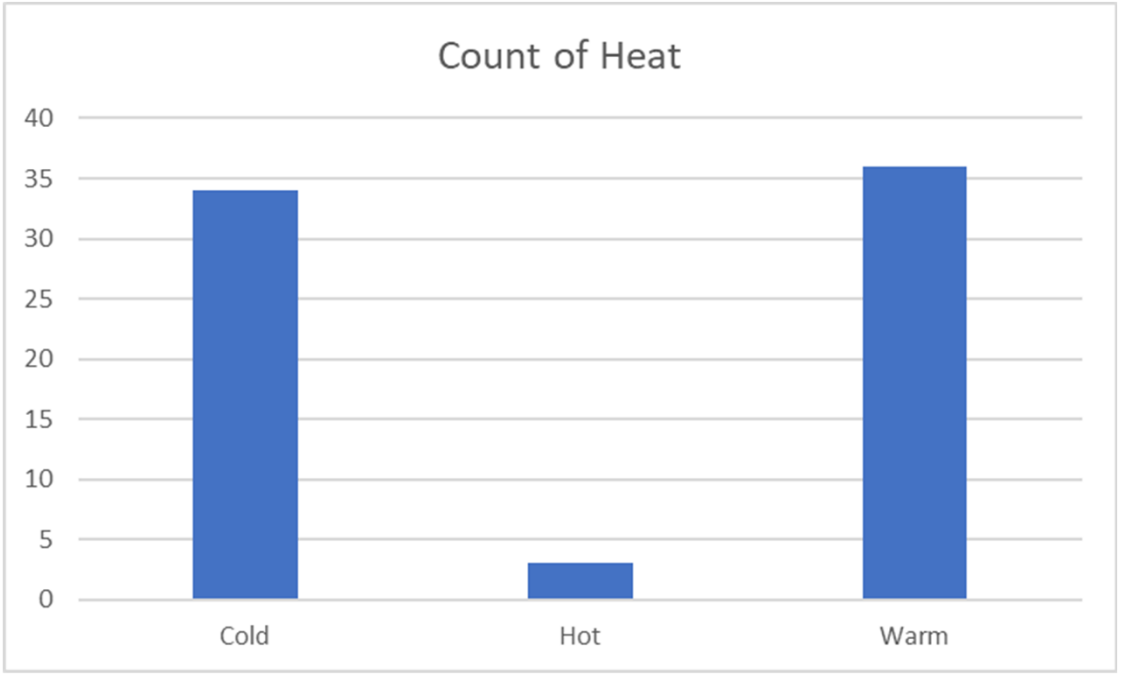
	Est Die	Flip Stock	Sales History	Total Quotes	Quoted Amount	Authorized Availability	Authorized Price
Missing Values	65	71	69	61	61	26	25
% Missing	89.0%	97.3%	94.5%	83.6%	83.6%	35.6%	34.2%

The objective of ETL stage is to ensure that there is no ambiguity in the data, such as “garbage” and missing values or outliers for numerical values. As the % missing values for these variables are relatively large, we excluded these variables from our analysis.

Data Audit – Source variables.

- As the percentage of missing values for those source variable is relatively large, we excluded these columns (source variables) from our analysis.
- The variables on which we perform our analysis are follow.
- **Family** - Non-ordinal categorical data. The family of electronic component.
- **Historic Customer** - Numerical data. The historical number of customers who have purchase this component between 2015-Present.
- **EOL_time** - Numerical data. Time (in years) since the component became obsolete.
- **Qty** - The number of components available in stock with the suppliers for sale.
- **Price** - The price at which the component is sold by the supplier.





Summary of Creating Analytical File

- The files we worked on had all the information in it, and we did not have to merge it with other files to create our analytical file.
- To simplify our classification analysis, and due to class imbalance problem, we relabelled all “Hot” labels as “Warm”.

Source Family	Source historic	Source EOL	Derived Time_EOL	Source Qty	Source Price	Target Heat
DSP56311	85	2018	2	1,967	34.76	Cold
MPC8241-D	51	2019	1	33,611	23.82	Warm
MPC8541_PQ37L	61	2019	1	2,612	147.1	Cold
MPC8555_PQ37L	49	2019	1	33	127.92	Warm
MPC862	30	2017	3	1,198	93.16	Warm

EOL_time was created as a derived variable from EOL (End of Life). Given as:
EOL_time = 2020 – EOL.
See Appendix for an example of analytical file with the categorical “Family” variable encoded.

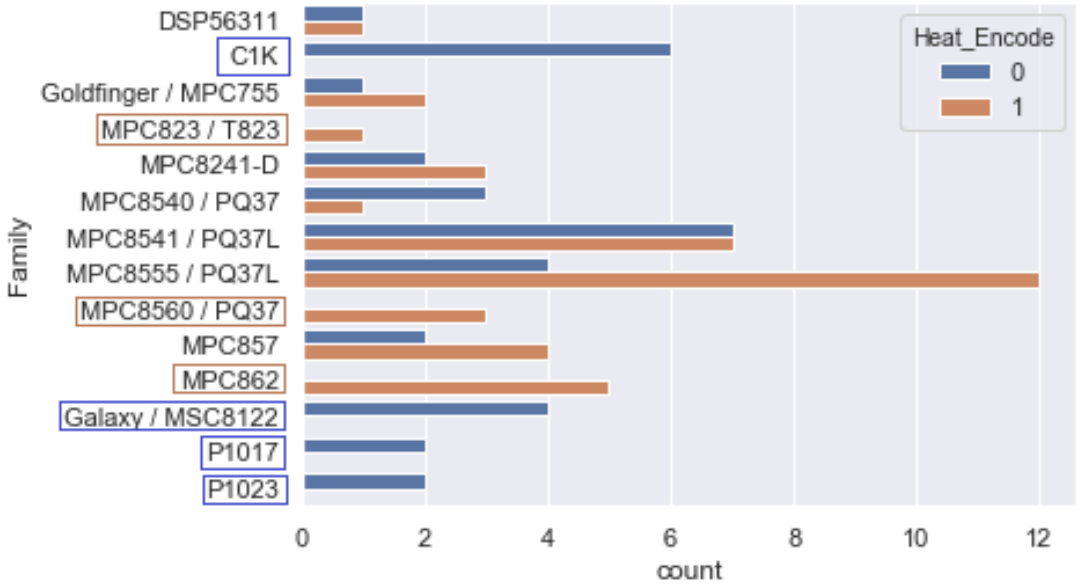
EXAMPLE OF AN ANALYTICAL FILE

Relative Importance of Variables – Correlation Analysis

Variable	Contribution %	Impact
Family	95.7	Variable - See Appendix
Qty	3.1	Positive
Price	1.2	Positive
Historic Customer	0	Positive
EOL	0	Negative

- So essentially this means that removing the "Historic Customer" and "EOL" from my predictive model, will have no impact on the performance of the predictive model.
- Removing "Price" in addition, will result in negligible degradation of performance.
- Both these statements have been verified by running predictive analysis after removing low-contributing variables.

Relative Importance of "Family" 0-Cold, 1-Warm

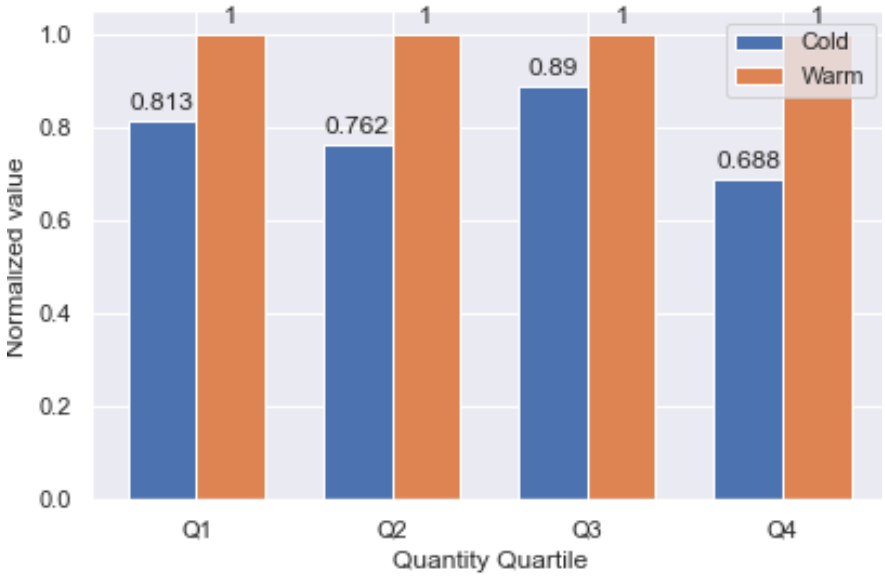


- ✓ It can be seen that all C1K, Galaxy / MSC8122, P1017 and P1023 components are exclusively "Cold".
- ✓ Whereas MPC823 / T823, MPC8560 / PQ37, and MPC862 are exclusively "Warm".

Relative Importance of "Quantity"

- **TREND:** We can see a clear trend that as for all the four quantiles of quantity, "cold" components are ordered in lesser quantity compared to the "warm" components.

- Normalized values are calculated as follow. For example in Q1, the average value of quantity is 36.25 for "cold" components, and the average value of quantity for "warm" component is 44.57. Both the values are divided by 44.57 to obtain the normalized values.

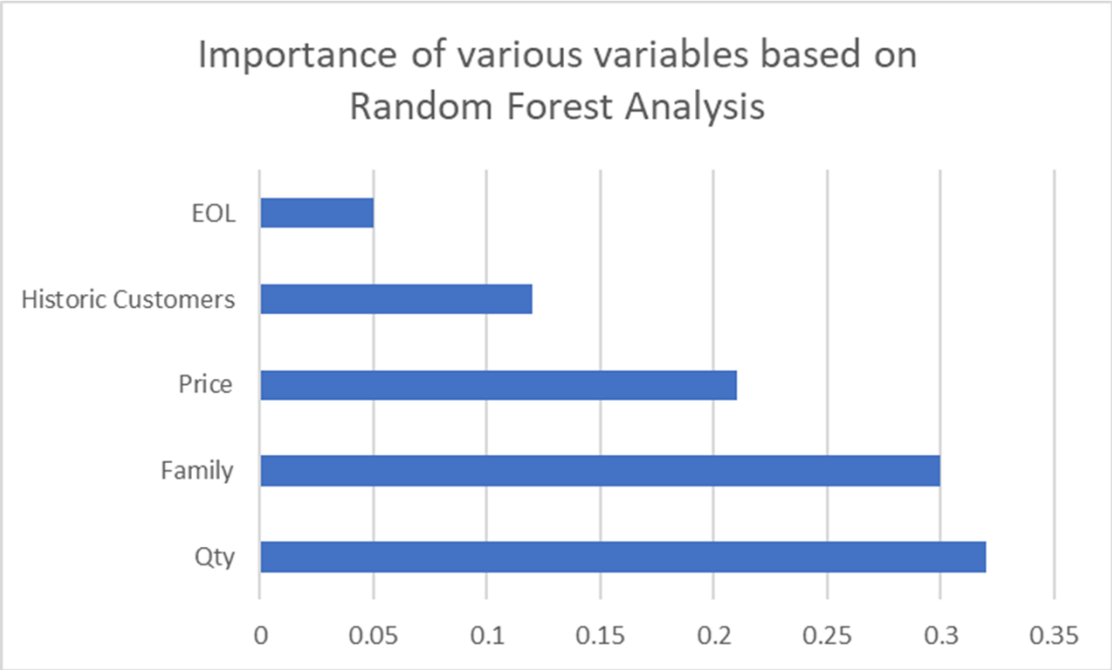


Quantity Value is increasing

Relative Importance of "Price"

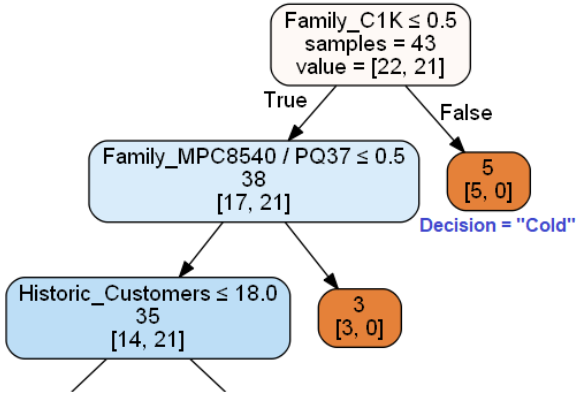
- TREND: It can be clearly seen that the likelihood of an electronic component being "Warm" increases as the price of the component increases.

Relative Importance of Variables – Forest Tree



This result validates our prior analysis that EOL and Historic Customers have minimal impact on the decision of the outcome.

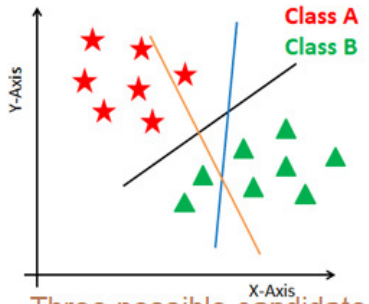
Classification Algorithms



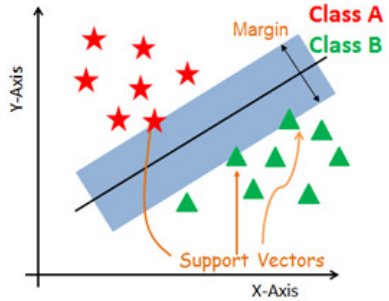
The decision tree classifiers make decision about the outcome based on a series of *useful questions*.

- We evaluated a variety of different classifiers and choose the best 5 for our Majority Voting Decision.

1. Decision Tree (illustrated)
2. Support Vector Machine (SVM)
3. Neural Network
4. Random Forest
5. Gradient Descent
6. Logisitic Regression
7. K – Nearest Neighbors

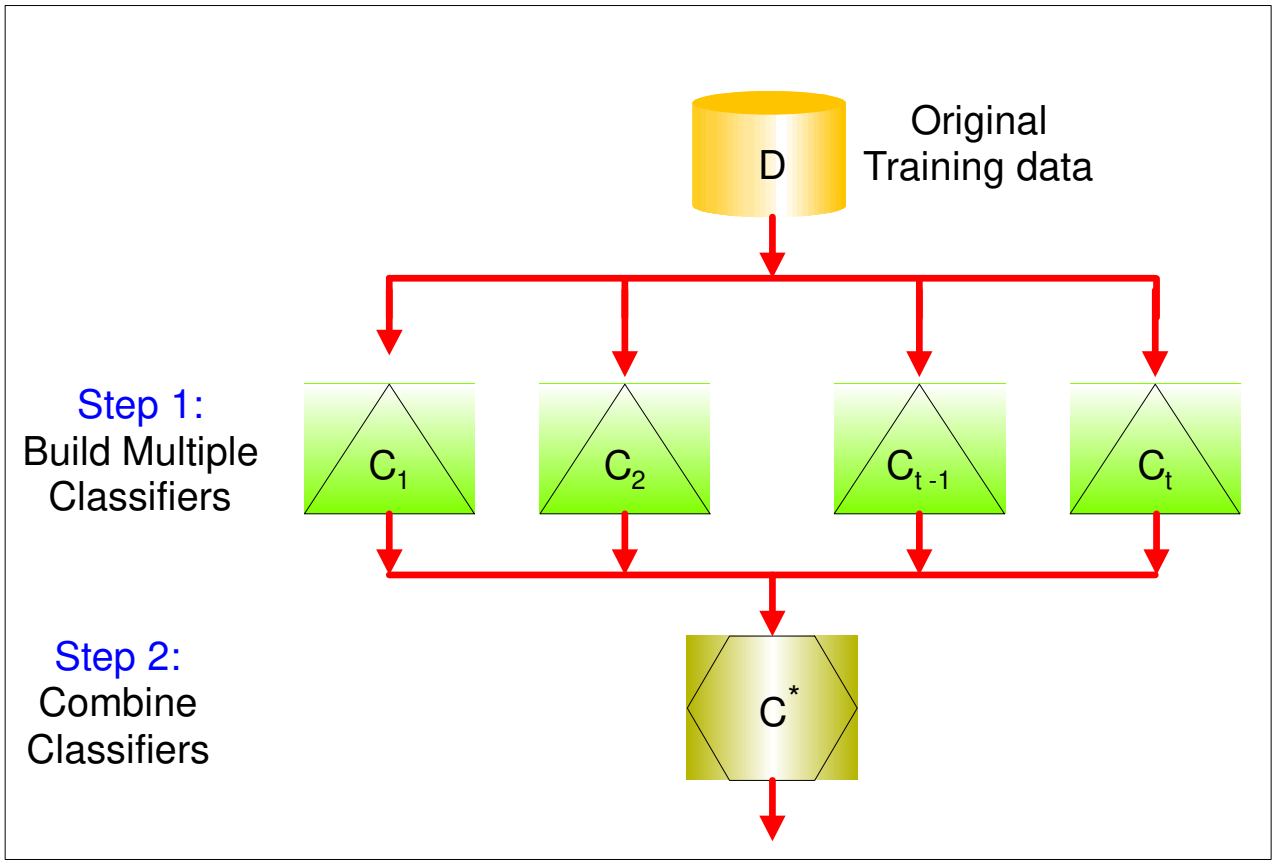


Three possible candidate for plane (line).



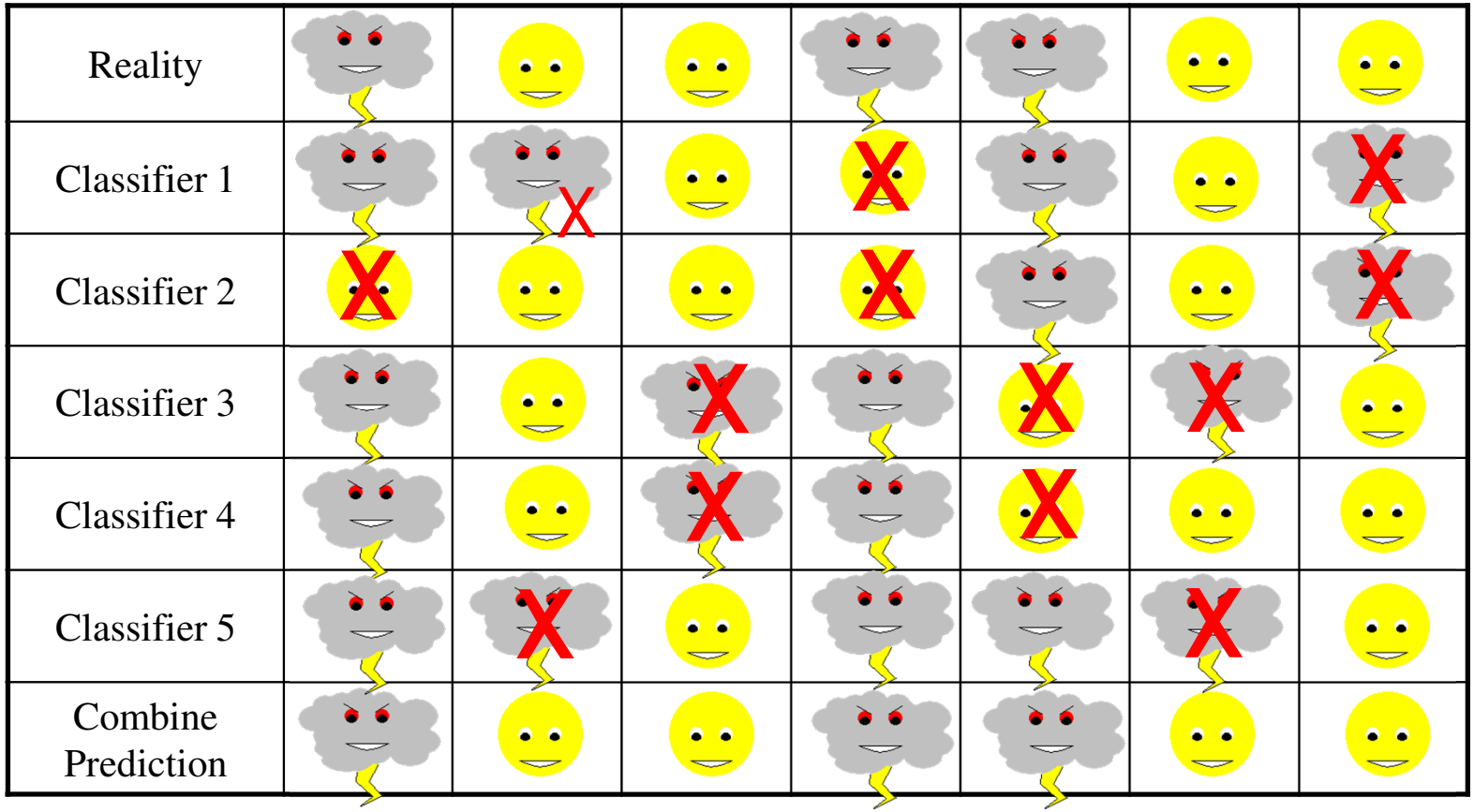
SVM selects the line which (i) sepataes the classes and (ii) maximize the margin

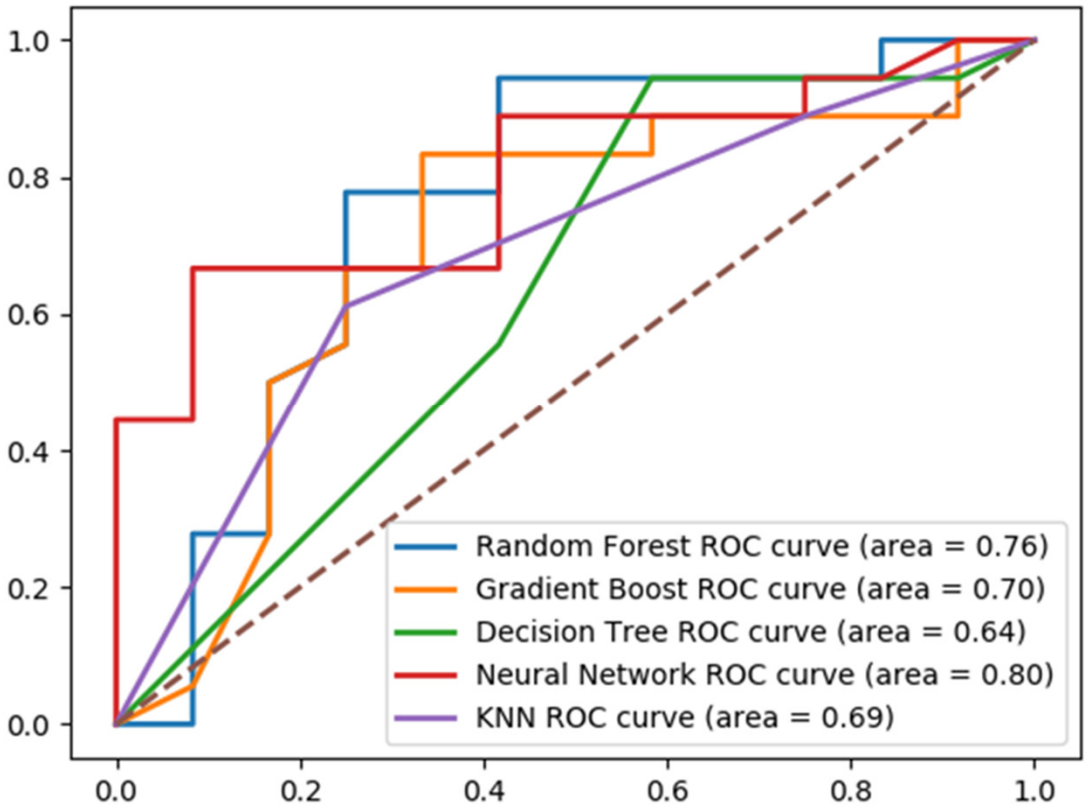
Proposed Model – Majority Vote



Example (for illustration):

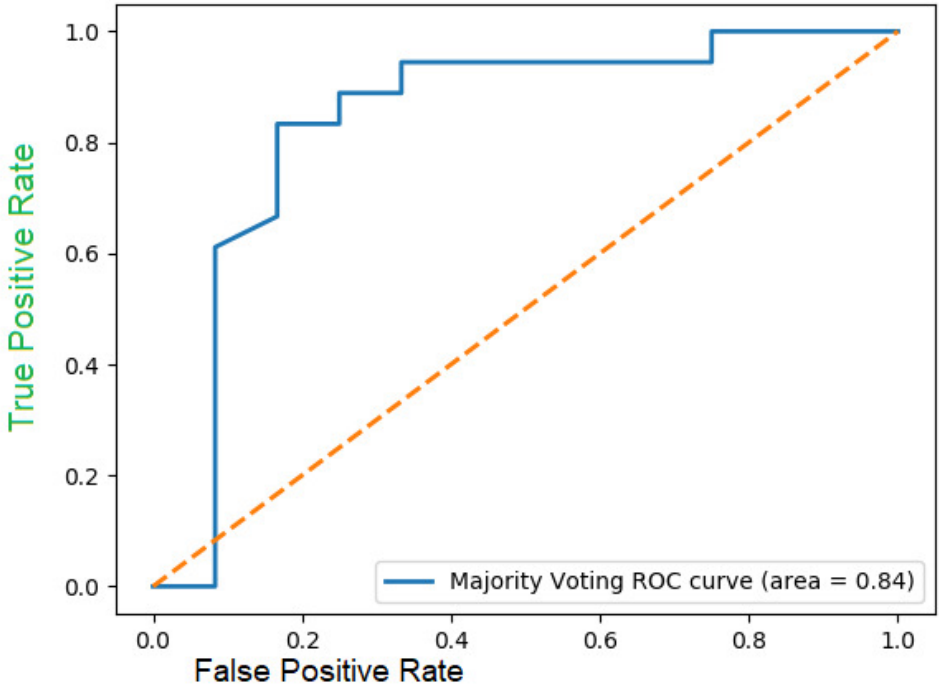
Weather Forecast. Each of the five classifiers has an accuracy of ~63%, but due to the combined voting of the all classifiers we obtain 100% accuracy.





AUC CURVE – TOP 5 CLASSIFIERS.

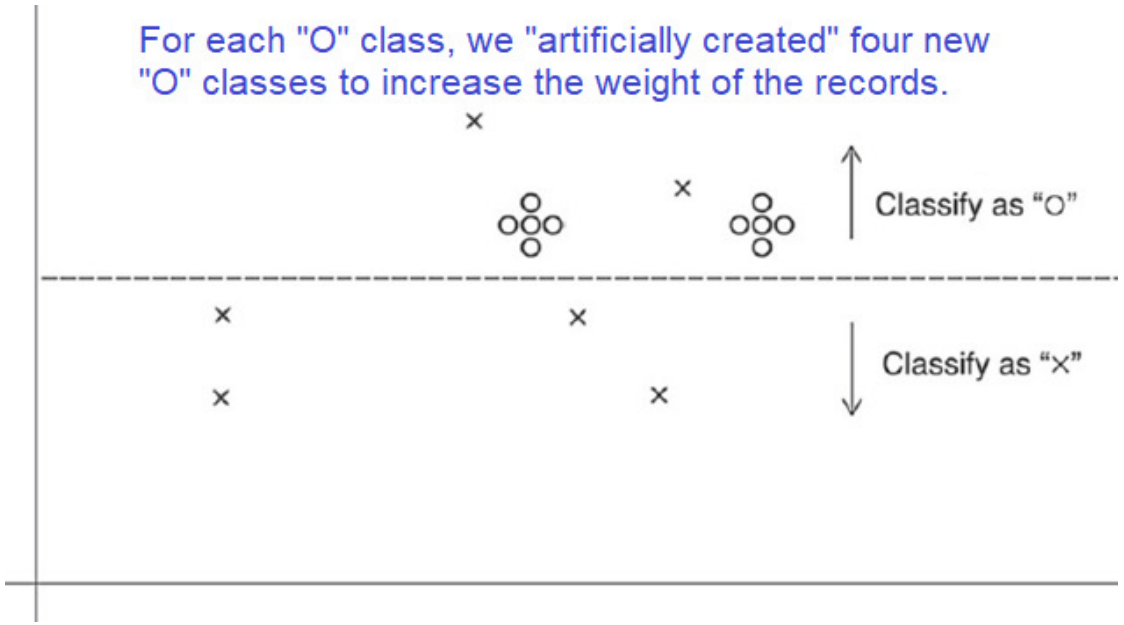
- Results shown after discarding low performing classifiers.
- Higher the area, better the performance.
- Dashed line show performance of a naïve randomized approach to make decision (e.g. classify all results as "Warm"), with an area of 0.5.



Accuracy of validation data:
Confusion Matrix (Accuracy 0.8000)

	Prediction	
Actual	Cold	Warm
Cold	8	4
Warm	2	16

AUC CURVE – MAJORITY VOTING.
HAS MUCH BETTER PERFORMANCE THAN EACH OF THE INDIVIDUAL CLASSIFIERS.



Confusion Matrix (Accuracy 0.6691)

	Prediction	
Actual	Cold	Warm
Cold	362	181
Warm	1	6

Renesas (Class Imbalance) data set (98.5% cold).

- Oversample rare records.
- As the business value of Warm records is much higher, we can not afford to misclassify those records.
- High Recall Rate. 6/7=86%.

Business Value of the Proposed Model

We have done an analysis of how various variables can influence the sale-ability of various electronic components.

Based on our analysis f the various variables, we proposed an ensemble learning based classification predictive model.

Our model can be extended to data set with class imbalance to achieve high Recall Rate, at tradeoff cost of relatively lower accuracy.

The results can facilitate Flip Electronics to decide which electronic components to procure, hence maximize pofit, by looking at various characteristics of the components.

Key Insight

Our proposed model produces a high accuracy rate of 80% on validation data.

For those records for which our model generated incorrect result – those were borderline case and did not lead to loss of business value.

Currently we are working on other data set with some variation on source variable, data set size and class imbalance.

As part of future work, the parameters of models such as Support Vector Machine and Neural Network can be optimized. And the results expanded for multi-classification.

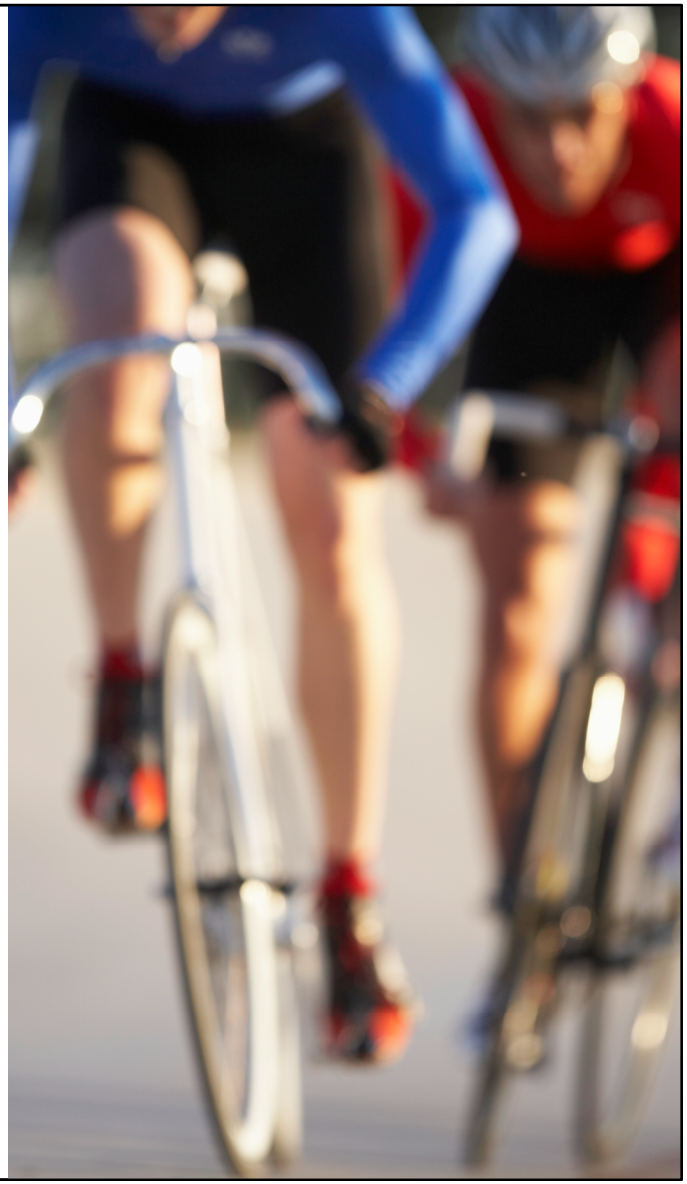
Recommendation – 1 Customer/ Market Segmentation

- By collecting more information about their customers such as company size, number of employee, annual turnover, industry, location, years since operations, the company can identify high yield future customers.

- Segments (cluster) are identified by managers based on their observation of the behavioral and “demographic” characteristics of likely users

Recommendation – 2
(High Worth) Customer Retention

- Flip can analyse the past sales (6-12 months) data about its customers and based on sales revenue during the first 6 months identify high worth customers who generated the most revenue for the company.
- A similar analysis should then be generated for the last 6 months, and then it should be evaluated whether the customer has switched to another distributor.
- This analysis can help identify whether revenue generation from such customers is decline, stable or growing.
- An increase marketing effort should then be made to those customers with declining sales revenue.



Recommendation 3 – Market Basket analysis

- Flip can perform a market basket analysis of purchase made by past customers, to analyze products which have association.
- For example, for the 5 transaction ids it can be seen that customers who buy Milk, are more likely to buy Diaper in the same transactions.
- This analysis can then help Flip to do targeted marketing of electronic component to their potential customers.

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

THANK YOU