



Final Project

GROUP 11

APALA MISHRA

DHWANI OJHA

DIVYA MARKANDEY

HASMITHA VASIREDDY

SHAGUN GUPTA

Contents

- ▶ The 3 Classification Problems
- ▶ Our Approach
- ▶ Vendors We Worked With
- ▶ Evaluation Criteria
- ▶ Evaluation
- ▶ Learnings

The 3 Classification Problems

The B Classifier

The B classifier is a simple classification problem, with data labeled “bo” or “bn” categories based on a single variable. It has only about 2000 rows in 1 file. The data values range from 0 to .110

The C Classifier

The C classifier uses 5 files (initially), and then 500 files, labeling the data into “cb” and “cw” categories based on a single variable. The data values range from 0 to slightly over 250.

The 3 Classification Problems

The O Classifier

The O classifier uses 1 file labeling the data into “o0”, “o1”, “o2”.... “o9” categories based on 8 variables. The data values range from 0.01 to to around 2.24.

Our Approach for our Classifiers

- ▶ Based B and C classifiers of visualizations initially, later switched to a Bayesian classifier and Logistic Regression respectively to make it more dynamic.
- ▶ O classifier had more complex values and more variables and visualizations were not conclusive, we examined the mean, median, and variance next and looked at box plots to understand these metrics. Again, it was hard to get conclusive findings.
- ▶ Finally, we tried SVM and Random Forest, and found that it gave us a high accuracy score.

Vendors Evaluated

Opti-Metric Solution

NextUp

Smart Sapiens

Code Crazies

Data Wizards

QuickTech Inc

Data Dynamos

Data Nest

Analytical Nexus

ANS3

Evaluation Criteria

Our Evaluation Criteria is based on

- ▶ Accuracy score (percentage of correctly classified rows) = $\text{Number of correct predictions} / \text{total accuracy}$.
- ▶ Errors the code generated and how smoothly it ran
- ▶ Responses, effective solutions and customer support provided by the vendors
- ▶ Based on this criteria, the vendor we will be choosing to work with NextUp
- ▶ Source of samples we tested vendor code on- our group's data, G11

Evaluation

- Evaluation based on accuracy

B classifier			C classifier			O Classifier		
Groups	Work Successfully?	Accuracy?	Groups	Work Successfully?	Accuracy?	Groups	Work Successfully?	Accuracy?
Data Nest	Yes	100	Opti-Metric Solutions	Yes	96.69	NextUp	Yes	100
NextUp	Yes	99.9	NextUp	Yes	92.9	Code Crazies	Yes	100
Smart Sapiens	Yes	99.9	Data Wizards	Yes	88.1	Opti-Metric Solutions	Yes	99.75
Opti-Metric Solutions	Yes	99.85	Analytical Nexus	Yes	88.1	Quick Tech Inc	Yes	99.75
Data Dynamos	Yes	99.85	Data Dynamos	Yes	79	Data Dynamos	Yes	99.7
Analytical Nexus	Yes	99.85	Quick Tech Inc	Yes	78.86	Smart Sapiens	Yes	99.5
ANS3	Yes	99.85	ANS3	Yes	78.6	Data Wizards	Yes	99.5
Data Wizards	Yes	99.7	Data Nest	Yes	78	Analytical Nexus	Yes	99
Quick Tech Inc	Yes	99.65	Smart Sapiens	Yes	75.87	ANS3	Yes	87.58
Code Crazies	Yes	80.6	Code Crazies	Yes	56	Data Nest	Yes	100
Average		97.915	Average		81.212	Average		98.478

Link to Excel: [FP-EVALS-FINAL.xlsx](#)

Learnings

- ▶ It is important to have a single point of contact, i.e, a project manager, who understands the business/technical problem at hand and can seamlessly connect between parties involved
- ▶ It is important to work with a limited number of vendors to focus on getting the best result. Working with too many vendors may take away from quality focus and improvisation. Of course, given the class size, the number of vendors could not be limited, but in the real world we believe that shortlisting a handful of vendors will work wonders in terms of getting the best output
- ▶ It is important to have a dynamic solution to the problem that will adapt to the change in data. For example, we initially had a classifier for week 7 and 8 that was based on visualizations of the data. We found a clear value that divided the classes, however, when more complex data is introduced or if the values change, this will not hold and lead to many misclassifications. This will also work on different datasets

Learnings

- ▶ It is important to have a test template to test multiple solutions (especially if you are working with multiple vendors), and provide the vendors with this template
- ▶ Important to share appropriate files, updated models and updated communication in real time. Challenge- too many cross-team reworkings to stay updated, feedback should ideally be from one source only
- ▶ Overall, a good experience to understand how communication and codes need to be streamlined, and the importance of a good structure in hand for optimal results



Thank You