MACHINE LEARNING TECHNIQUES FOR THE DETECTION OF UNFAIR PRICING IN SUPERMARKETS ACROSS TRINIDAD AND TOBAGO

Arti K. Ramdhanie*

Faculty of Science and Technology, The University of the West Indies, Trinidad
Email: arti.ramdhanie@sta.uwi.edu (Corresponding author)

Abstract: The tracking of prices in monitored supermarkets across Trinidad and Tobago is done by the Ministry of Trade and Industry. This initiative involves data collection every month for 118 grocery items ("standard basket"). The task of identifying which supermarkets are non-conforming in their pricing schemes is linked to the "total basket price" (total cost of the 118 items). An outlier is defined as any datapoint that varies significantly from all other observations in a dataset. In this paper, it is any supermarket that exceeds this total basket price by 5%. The aim of this research was twofold, with the first goal being to employ feature selection methods to reduce the number of items being collected. The second goal was to create a logistic regression learning model that can identify whether supermarkets are non-conforming, given their pricing information. The dataset contained 692 datapoints and out of these, only eight (8) were classified as outliers. This is an imbalanced dataset. Resampling by SMOTE (Synthetic Minority Oversampling Technique) was used to synthetically generate data for the training set. Seven (7) feature selection methods were also investigated and their results discussed and analysed. In doing this, a more balanced dataset was achieved which was tested and validated on the unseen data (testing set). The metrics indicated that a subset of these features can be collected whilst still maintaining the supermarket outliers.

Keywords: *Feature Selection, Logistic Regression, Machine Learning, Outlier Detection, SMOTE.*

https://doi.org/10.47412/GIDS9258

1. Introduction

Monitoring the price of items in Trinidad and Tobago has proven to be an important task given the current state of the economy. Prices for standard goods are tracked in order to provide this information to the public. This allows consumers to potentially avoid supermarkets which are unfair in their pricing methods. For the detection of these non-conforming supermarkets, a machine learning technique based on binary classification is proposed. Several feature selection methods are also suggested with the aim of reducing the amount of man power required in the collection of these prices.

2. Dataset Information

2.1 Source

The data was accessed from the Ministry of Trade and Industry website under their Publications section [1]. Every month, a .pdf or .xlsx file is generated with the pricing information for the "standard basket".

2.2 Data Processing

Prices are collected by months so each file was loaded, data cleaned, pivoted, encoded, and merged to create one (1) master table with the structure as seen in Table 1.

Table 4:Dataset Information

Column Name	Description
Grocery	Name of Supermarket
Area	Geographic Area
Date	Date of collection (mm-YY)
v1 – v118	Pricing information for 118 items

2.3 Outliers

An outlier is defined as a datapoint in a dataset that varies significantly from all other observations. It lies outside the general pattern of distribution [2]. For this research, an outlier in this dataset is defined as one whose total basket price exceeds more than 5% of the average basket price for that month – see Algorithm 1. An 'Outlier' column was added to the dataset. If a supermarket was an outlier, a '1' would be the Outlier Value as opposed to a '0' which represents a non-outlier supermarket. The results of Algorithm 1 indicated that the distribution of outliers to non-outliers was highly imbalanced (Table 2). A dataset is identified as imbalanced if the classification targets are not equally represented. For example, if there are two (2) values for the target variable, a balanced dataset would mean that the distribution for each class is approximately 50%. Traditional machine learning methods are biased towards the majority class. This bias occurs because machine learning algorithms are usually evaluated using accuracy but imbalanced classes can actually have a high accuracy without actually making any useful predictions.

ALGORITHM 1: OUTLIER DETECTION ALGORITHM

```
Algorithm OutlierDetection(T[i], K) {
For all i:
        Average_Price = K
        Buffer_Val = 0.05 * Avg_Price;
        Outlier_Val = Avg_Price + Buffer_Val;
        If T[i] > Outlier Val
        Outlier = 1
        Else
        Outlier = 0
}
```

Algorithm 1: Outlier Detection

Table 5: Outlier Value Breakdown

Outlier Value	Count	%
0	684	98.84
1	8	1.16

2.4 Resampling Methods

Resampling is done in order to obtain (approximately) the same number of instances for both the target classes. Oversampling by Synthetic Minority Over-sampling Technique (SMOTE) was implemented. SMOTE creates synthetic observations based on the existing minority class utilizing Algorithm 2 [3]. The dataset was split into the Training and Testing subset and SMOTE was then applied to the Training Set alone (Table 3). This was done because if SMOTE is applied to the full dataset then synthetic data will closely resemble data in the testing subset. This can lead to a higher accuracy. Whereas if we split the data and then perform SMOTE on the training subset alone, the model can be evaluated using real, new data.

ALGORITHM 2: SMOTE

I. Consider a minority class sample
II. Compute its 'k' nearest neighbors
III. Pick one k point at random
IV. For each feature in the feature space:
    Compute the difference between the two points in terms of that feature
    For the synthetic datapoint, define the value of this feature as a random combination of the two points.
V. Repeat for each sample point until the desired level of balance is achieved

Algorithm 2: SMOTE

Table 6: Training Set SMOTE Breakdown

Outlier Value	Count	%
0	183	53.35
1	160	46.65

3. Feature Selection

Feature selection is the technique of extracting a subset of relevant features from the full dataset. It is defined as a process that chooses a minimum subset of M features from the original set of N features [4]. In this dataset, we have input $X = \{x_1, x_2,\ldots,x_N\}$ (where N = 118) and an output Y which contains our target variable. In most cases, the output Y is not determined by the complete set of features but some subset $\{x_1, x_2,\ldots,x_n\}$, where n < N. Thus, we end up with a set $T = \{(x_m, y_m) \mid m = 1...M\}$ of M training samples. In this case we have 692 training samples. For every x_m, we have a corresponding y_m target label. The feature selection methods used are: Manual Variance Threshold, Random Forest, Chi-Squared Filter, Entropy Based Filters, And OneR.

3.1 Manual Variance Threshold

The variance of each of the 118 items was calculated and a decision was made on whether to keep or discard the feature (Algorithm 3). Using Algorithm 3, the dataset was reduced from 118 features to 95 features. This is a reduction of 19.49% in features.

Algorithm 3: Variance Threshold Method	
I.	Calculate variance of item i (i = 1 to 118).
II.	Store in vector V = {$v_1, v_2,\ldots,v_N$}.
III.	Sort V in ascending order.
IV.	Remove feature with the lowest variance from $X = \{x_1, x_2,\ldots,x_N\}$.
V.	Check outliers.
VI.	If outliers remain the same, remove next feature with the lowest variance.
VII.	Repeat until outliers differ from original dataset.

Algorithm 3: Variance Threshold

3.2 Random Forest

This method finds the weights of attributes using the Random Forest Algorithm. It is an ensemble classifier utilizing numerous decision trees where each tree is a sequence of 0/1 questions based on a single/combination of features. At each node of the tree, the dataset is divided into two (2) categories with similar observations [5]. A vector of importance is generated based on the Random Forest Algorithm and the bottom 25% is selected.

3.3 Chi-Squared Filter

This algorithm finds weights of discrete attributes based on a chi-squared test. It is used to test the relationship of our feature variables, $X = \{x_1, x_2,\ldots,x_N\}$ against our target variable and gives us the Cramer's V coefficient. The vector of importance is generated and the bottom 25% is selected.

3.4 Entropy Based Filters

The algorithms in this section find feature importance values based on their correlation with the class attribute. Entropy is defined as the average rate at which information is produced by the dataset [5]. The function for entropy is given by Eq. (1).

$$H(X) = -\sum_{i=1}^{n} p_i * \log_b(p_i) \tag{1}$$

where
n = number of different outcomes (2 – class 1 and class 0)
p_i = probability of the ith class
b = base of the logarithm used

Entropy formula for this dataset is therefore given by Eq. (2).

$$H(X) = \left[-p_{(class\ 0)} * \log_2 p_{(class\ 0)} \right] - \left[p_{(class\ 1)} * \log_2 p_{(class\ 1)} \right] \tag{2}$$

3.4.1 Information Gain (Entropy Based Filter 1)

This measures how much information a feature provides about the target attribute. It can detect the feature(s) possessing the most information, based on a specific class. It is calculated using Eq. (3).

$$Information\ Gain = H(Class) + H(Attribute) - H(Class|Attribute) \tag{3}$$

3.4.2 Gain Ratio (Entropy Based Filter 2)

This is a modification of the Information Gain method above that reduces the bias by taking the intrinsic information into account. It can be calculated using Eq. (4).

$$Gain\ Ratio = \frac{H(Class)+H(Attribute)-H(Class|Attribute)}{H(Attribute)} \quad (4)$$

3.4.3 Symmetrical Uncertainty (Entropy Based Filter 3)

This is also a modification of the Information Gain method. It is normalized and adjusted for bias correction. It is calculated using Eq. (5).

$$Symmetrical\ Uncertainty = \frac{H(Class)+H(Attribute)-H(Class|Attribute)}{H(Attribute)+H(Class)} \quad (5)$$

3.5 OneR

This algorithm uses the OneR classifier to calculate a feature's weights. For each individual attribute and its value, the error produced is the error if only that feature was used in the classifier [6]. It separates the series of values into many disjoint intervals and evaluates the features according to their error rates. A vector of feature importance is also generated by this algorithm.

4. Subset Correlation

The seven (7) subsets were analyzed and the features that were common to four (4) or more were regarded as the "top features for removal". The Coefficient of Variation (CoV) of these items were then calculated. This is a statistical measure of the dispersion of datapoints around the mean. The advantage of using the coefficient of variation is that it does not have a unit of measurement so it will be universal across all datasets. The formula is given by Eq. (6).

$$Coefficient\ of\ Variation = \frac{\sigma}{\mu} \quad (6)$$

where
σ = standard deviation
μ = mean

The higher the coefficient of variation value, the greater level of price dispersion is present for that feature. The items in Table 4 all have relatively low values. This means that their prices had little volatility across all supermarkets and reasonably did not contribute to the supermarket being an outlier. In Table 4, we see that four (4) items belong to six (6) or more subsets.

These items (and their corresponding coefficient of variation values) are:

- v42 – Pigeon Peas (Pre-Packaged) 400g – 0.0257
- v107 – Sugar (Loose) 454g – 0.0867
- v40 – Split Peas (Pre-Packaged) 400g – 0.1187
- v46 – Curry Powder 85g – 0.0664

These items, for example, are all purchased from local suppliers which explains why the prices do not exhibit any volatility. These items will normally come with a "suggested market price" from the providers. The Pigeon Peas (v42) is the only item to belong to all seven (7) feature selection subsets. After investigations, it was found that this item is distributed by Pepe's Marketing in St. Augustine, Trinidad and their suggested retail price is $8.00TTD. The prices in our dataset for v42 ranged from $8.00TTD to $11.00TTD.

The item in our dataset with the highest coefficient of variation is v91 (CORN FLAKES (Local) 200g). This item is distributed by Sunshine Snacks which is a subsidiary of Associated Brands Industries Limited. The price for this item ranged from $10.00TTD to $110.90TTD. There were three (3) supermarkets that charged higher than average prices for this product and they were all located in Tobago. The reason for this type of price gouging is unknown. Another example of these disparate pricing practices would be the items with the 2nd and 3rd highest coefficient of variations. Their ranges were [$3.00TTD, $28.00TTD] and [$1.82TTD, $27.27TTD] respectively. It is interesting to note that all of the extreme prices above were recorded in Tobago.

Table 7: Features with subsets in common

4 Subsets in Common	**5 Subsets in Common**	**6 Subsets in Common**	**7 Subsets in Common**
v12, v14, v27, v36, v4, v47, v61	v101, v111, v20, v30, v37, v58, v63, v70, v76, v77, v78, v81, v84, v86, v87, v90	v107, v40, v46	v42
Total = 7	Total = 16	Total = 3	Total = 1

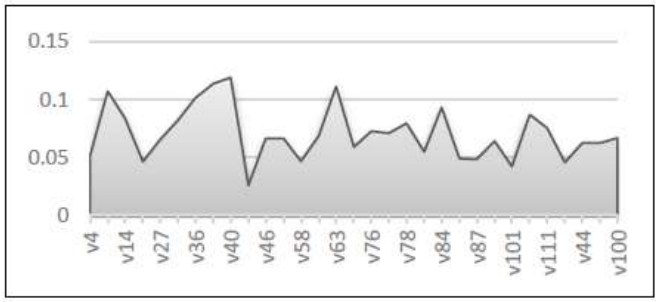

Figure 20: CoV for Feature Removal

5. Logistic Regression

This model belongs to the family of Generalized Linear Models (GLM) and it is a binary classification algorithm which is used when the target response is dichotomous (1/0, "yes"/""no"). GLMs comprise a linear combination of input features and the mean of the response is related to these features by a link function [7]. Equation (7) follows a sigmoid function which limits the range of the probabilities between 0 and 1.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \tag{7}$$

where

p = P (Y=1 | X)
$\beta_0 + \beta_1$ = linear combinations of independent variables
$\log \frac{p}{1-p}$ = log(odds) or logit function

The method used to fit the logistic regression model is achieved through Maximum Likelihood Estimation. It uses this to find the best coefficients such that the predicted probabilities are as close to the observed probabilities as possible. The likelihood function is denoted by Eq. (8).

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i} * (1 - p_i)^{1-y_i} \quad (8)$$

The Bayesian Generalized Linear Model was used as opposed to GLMs in order to combat the problem of separation. This occurs when there are dichotomous outcomes in the target class (in this case, 0/1). Separation therefore happens if some linear combination of our predictor variables, X = {x_1, x_2,….,x_N}, are associated with only one outcome value when the predictor is greater than some constant value [7]. The Bayes GLM is a modification to the standard GLM that uses an approximate EM algorithm to update the beta coefficients at each step using an augmented regression to represent the prior information [8]. The Student-t prior distributions are used for the coefficients.

Features were eliminated in increments of 5% up to 25%. After each increment, the logistic model was run and the results recorded in Table 5-11.

Table 8: Variance Threshold Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9398	1.0000	0.9391	0.1600	0.2759
15%	0.9398	1.0000	0.9391	0.1600	0.2759
20%	0.9398	1.0000	0.9391	0.1600	0.2759
25%	0.9398	1.0000	0.9391	0.1600	0.2759

Table 9: Random Forest Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9398	1.0000	0.9391	0.1600	0.2759
15%	0.9456	1.0000	0.9449	0.1739	0.2963
20%	0.9427	1.0000	0.9420	0.1667	0.2857
25%	0.9456	1.0000	0.9449	0.1739	0.2963

Table 10: Chi-Squared Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9427	1.0000	0.9420	0.1667	0.2857
15%	0.9427	1.0000	0.9420	0.1667	0.2857
20%	0.9513	1.0000	0.9507	0.1905	0.3200
25%	0.9542	0.75	0.9565	0.1667	0.2727

Table 11: Information Gain Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9398	1.0000	0.9420	0.1304	0.2222
15%	0.9427	1.0000	0.9420	0.1667	0.2857
20%	0.9542	1.0000	0.9536	0.2000	0.3333
25%	0.9427	1.0000	0.9420	0.1667	0.2857

Table 12: Gain Ratio Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9341	0.7500	0.9362	0.1200	0.2069
10%	0.9341	0.7500	0.9362	0.1200	0.2069
15%	0.9398	1.0000	0.9391	0.1600	0.2759
20%	0.9370	1.0000	0.9362	0.1538	0.2667
25%	0.9456	1.0000	0.9449	0.1739	0.2963

Table 13: Symmetrical Uncertainty Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9398	1.0000	0.9391	0.1600	0.2759
15%	0.9398	1.0000	0.9391	0.1600	0.2759
20%	0.9456	1.0000	0.9449	0.1739	0.2963
25%	0.9513	1.0000	0.9507	0.1905	0.3200

Table 14: OneR Evaluation Metrics

%	Accuracy	Sensitivity	Specificity	Precision	F1 Score
5%	0.9370	1.0000	0.9362	0.1538	0.2667
10%	0.9370	1.0000	0.9362	0.1538	0.2667
15%	0.9398	1.0000	0.9391	0.1600	0.2759
20%	0.9398	1.0000	0.9391	0.1600	0.2759
25%	0.9370	1.0000	0.9362	0.1538	0.2667

5.1 Evaluation Metrics

I. Accuracy: This is defined as the proportion of correctly classified data points and it is the most common metric used for evaluation.
II. Sensitivity: This calculates the proportion of positives identified correctly by the classifier.
III. Specificity: This calculates the proportion of negatives accurately identified by the classifier.
IV. Precision: This calculates how much of the classified/predicted positives are actually positive.
V. F1 Score: This is also considered a test of the classifier's overall performance. It is the harmonic mean between the Precision and Recall.

For the purpose of this research, the important metric to note is Sensitivity.

6. Complete Analysis

Table 12 shows the seven (7) feature selection methods utilized in this paper and their best subset percentages. Information Gain, with a 20% subset removal, provides the overall best F1 Score. Symmetrical Uncertainty gave the second best F1 Score with an elimination of 25% of the items. The available dataset has 118 unique features so a removal of 25% translates to 30 items being removed as opposed to 20% which eliminates 24 items.

Table 15: Final Results

Feature Selection Method	%	Sensitivity	F1 Score
Variance Threshold	25%	1.0000	0.2759
Random Forest	25%	1.0000	0.2963
Chi Square Filter	20%	1.0000	0.3200
Information Gain	20%	1.0000	0.3333
Gain Ratio	25%	1.0000	0.2963
Symmetrical Uncertainty	25%	1.0000	0.3200
OneR	20%	1.0000	0.2759

7. Conclusion

An outlier in this paper is any supermarket whose total basket price is more than 5% above the average basket price. This indicates that the supermarket in question is inflating their prices. In the supermarket-price dataset, there was a total of eight (8) outlier supermarkets – a 1.16% compared to the 98.84% of non-outlier occurrences. The task of handling imbalanced datasets is proving to be very common in real life.

Synthetic Minority Over-sampling Technique (SMOTE) was used to tackle the class imbalance by creating synthetic observations based on the existing minority class. After the SMOTE algorithm was applied, the balanced dataset became 53.35% non-outliers and 46.65% outliers.

The first aim of this research was to reduce the number of items being collected. Seven (7) feature selection techniques were investigated. They were: Manual Variance Threshold, Random Forest, Chi Square Filter, Information Gain, Gain Ratio, Symmetrical Uncertainty, and OneR. These algorithms ranked the importance of each attribute (product) and the bottom 25% was removed. The seven (7) subsets were analyzed and the features common to four (4) or more were identified as the "top features for removal."

The second aim of this research was to build a logistic regression model and test it on the seven (7) subsets generated by the feature selection algorithms. Features were removed in increments of 5% and the model was run on each increment. Comparing these results, shows that the Information Gain algorithm with a 20% deduction performed the best. The second-best performance was by the Symmetrical Uncertainty algorithm with a 25% reduction in features.

In conclusion, the two (2) aims outlined in this research were achieved. It was shown that we can potentially remove up to 25% of the total number of items being monitored (approximately 30 items) without affecting the ability to detect the non-conforming supermarkets.

References

[1] Ministry of Trade, Industry, (accessed July 20th, 2019). *https://tradeind.gov.tt/*
[2] M. Sugiyama, 2016. *Introduction to Statistical Machine Learning.* Morgan Kaufmann Publishers.
[3] T. Peters, 2018. *Detecting Credit Card Fraud Based on Transaction Data by Using a Novel Bump Hunting Method.* Universiteit Utrecht Netherlands.
[4] I. Guyon, A. Elisseeff. 2003. An Introduction to Variable and Feature Selection. In *Journal of Machine Learning Research 3* (pp. 1157-1182). JMLR.
[5] T. Hastie, J. Friedman, R. Tisbshirani, 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
[6] P. Chelvan, M. Ohana, K. Perumal. On Feature Selection Algorithms and Feature Selection Stability Measures: A Comparative Analysis. *International Journal of Computer Science & Information Technology* 9, no. 3 (2017).
[7] S. Menard. 2010. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE.
[8] A. Gelman, J. A. Jakulin, M. Pittau, Y. Su. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics* 2, no. 4 (2008).