TRANSACTION SUCCESS PREDICTION

CAN WE PREDICT FUTURE SUCCESSFUL TRANSACTIONS BASED PAST TRANSACTION DATA?

ROBERT S. BALCH II

USE CASES

PREDICTING FUTURE TRANSACTIONS CAN ALLOW FOR

- TARGETED MARKETING CAMPAIGNS
- MORE EFFICIENT RESOURCE ALLOCATION
- IMPORTANTLY - POTENTIALLY INCREASED REVENUE.

HOW THIS WAS ACHIEVED

- importing the data
- Data Cleaning and Exploration
- Correlation Analysis
- Data Splitting
- Model Training and Evaluation Using Naive Bayes Theorem
- Addressing Class Imbalances

ASSESSING THE DATA

STEP 1

Importing the data

The analysis begins with our dataset being pulled from "Transaction.training.csv"
Pandas was used to create the original DataFrame from the dataset and for visualizing the rows and columns through histogram plotting.

Visual analysis of column data helped identified the "target" column as our target data type.
This column's initial visualization indicated "successful sales = 1" and "unsuccessful sales = 0"
giving us our target to correlate predictors against.

STEP 2

Data Cleaning & Exploration

The dataset was found to have 53 columns and with a total of 180,000 rows approx.
Upon visual examination two columns whose data was not useful for our analysis were dropped - "Unnamed: 0" - showing a numbered list of column position/row in the table and "ID_code" - holding ticket string data whose content would not be viable for this analysis.
These columns were dropped leaving 51 columns by 180,000 rows approx.

STEP 3

Correlation Analysis

A correlation matrix was created to identify relationships between predictor columns and the target.
The data showed very low correlation scores with the highest coefficient values around .08

STEP 4

Data Splitting

The datasets were then divided into two DataFrames.
The first held successful "target" rows and the second held unsuccsessful "target" rows representing our successful transaction and unsuccessful transaction data.
The result was 161,960 "no_sales" rows and the remainder 18,040 "sales" rows.

STEP 5

Model Training and Evaluation using Naive Bayes Theorem

For training two more DataFrames were created.

20% of the data set was chosen to be set aside.
The model, trained over 50 iterations, achieved a mean accuracy of 91.11% on the original dataset, primarily reflecting its ability to predict unsuccessful transactions.

RESULTS

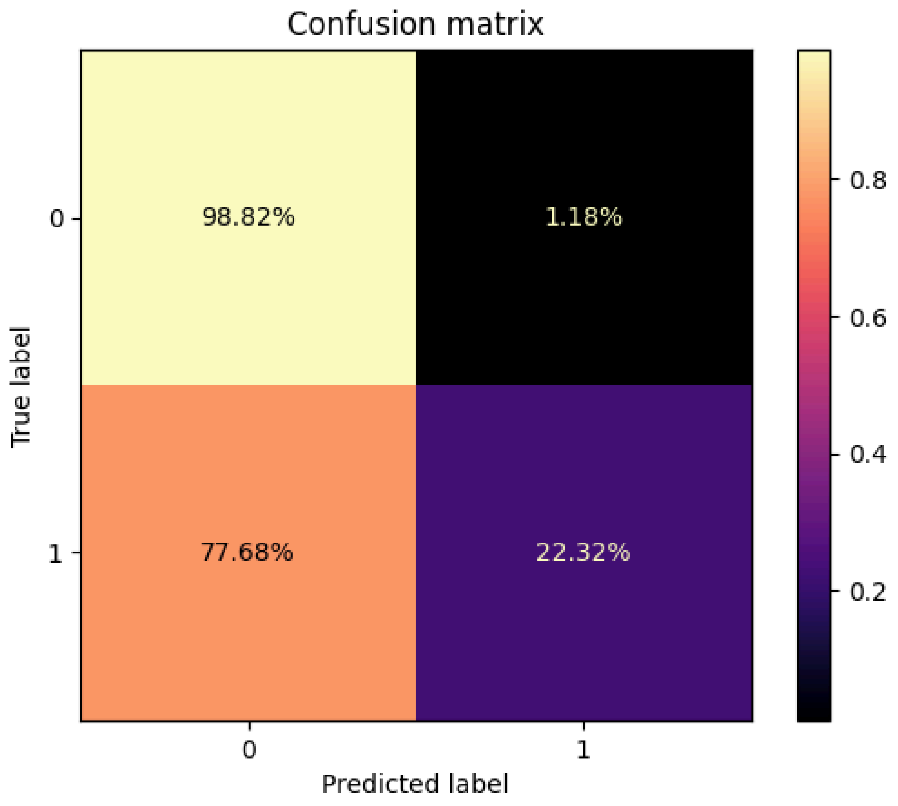
Addressing Class Imbalances

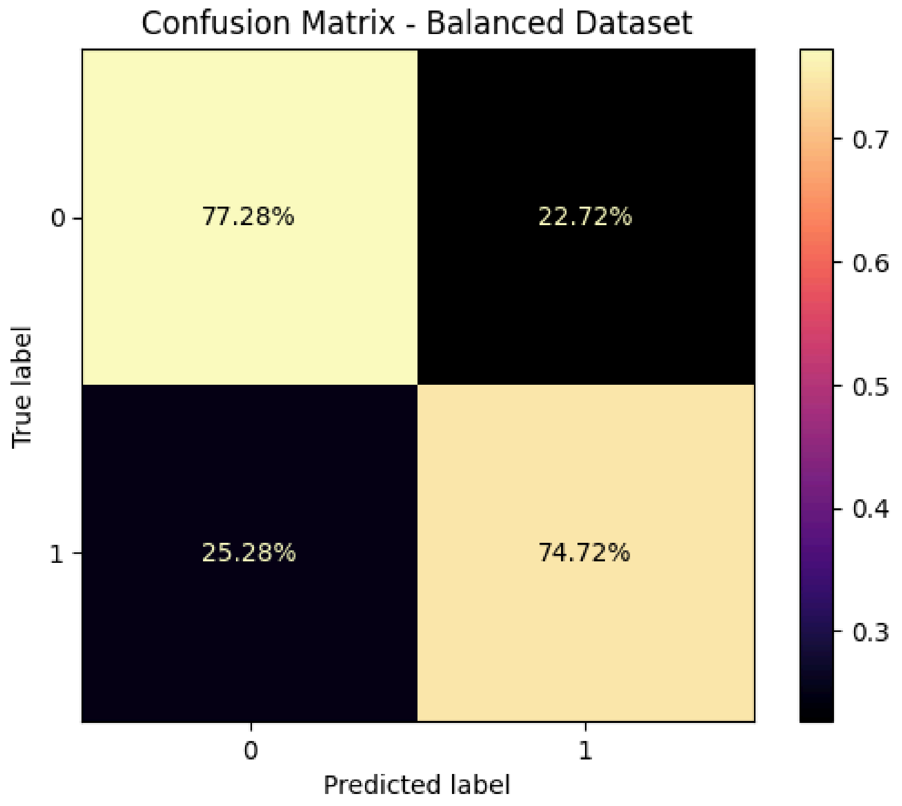
The analysis identifies a significant class imbalance, with a majority of transactions being unsuccessful. This imbalance biases the model, resulting in a high accuracy for predicting unsuccessful transactions but lower accuracy for successful ones. To address this the balanced model was created by randomly sampling an equal number of successful and unsuccessful sales variables.

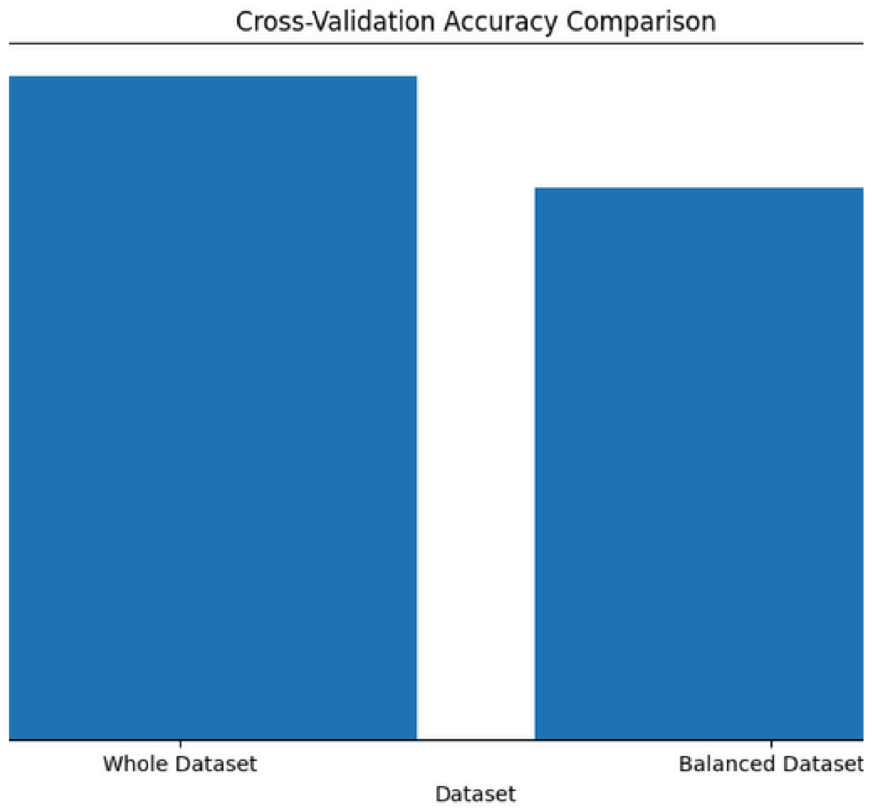
EVALUATING THE BALANCED DATA MODEL RESULTS

In the initial data set our confusion matrix was highly biased towards unsuccessful sales and showed inconsistency with our model prediction accuracy.

After balancing the dataset, the confusion matrix showed improved prediction accuracy for successful transactions, increasing from 1.18% to 22.72%. This indicates a more balanced model performance. Cross-validation results demonstrate improved consistency in model performance with the balanced dataset compared to the initial dataset, highlighting the impact of addressing class imbalance.







THE BOTTOM LINE?

While the model predicts possibly higher sales potential in the data when using the balanced dataset we also see some correlation with certain unknown var_* predictors. In particular some predictors correlate relatively strong with successful sales > .05 coefficients.

The variables var_20, var_31, and var_5 exhibit the highest correlation coefficients with successful transactions, ranging from approximately 0.07 to 0.08.

These correlations suggest a successful sale correlates with specific var_* predictors. While these correlation coefficients are quite small this suggests that focusing on these variables could potentially improve predictve models and sales strategies.
Keeping in mind that correlation is not causation and the dataset is highly skewed towards unsuccessful sales data is important in this regard and further analysis is required to determine the strength of these suggested correlations. Yet with this modelling we have a good starting point for future analysis.

- Importing the data
- Data Cleaning and Exploration
- Correlation Analysis
- Data Splitting
- Model Training and Evalusation Using Naive Bayes Theorem
- Addressing Class Imbalances
- Evaluating our models
- Considering the Bottom Line

THANK YOU!