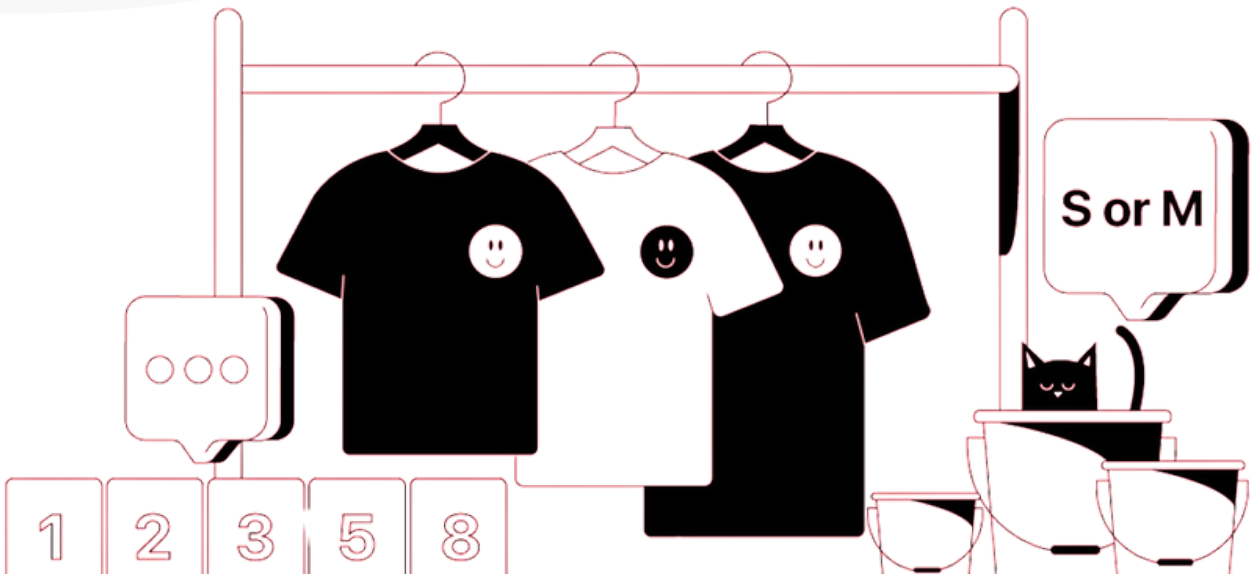
INVENTORY ANALYSIS BASED ON CLOTHING DESIGN SPECIFICATIONS

A case study in T shirt designs on amazon.in

BY
Naveen Kumar S
Jayasurya V

CONTENT



INTRODUCTION

01. PROBLEM STATEMENT

Optimize T-shirt inventory with data-driven insights to prevent overstocking and stockouts.

02. INDUSTRY APPLICATION

Optimize inventory by predicting demand and trends across all sales channels.

03.INDUSTRY (BUSINESS)

Data-driven inventory boosts efficiency, cuts costs, and enhances satisfaction.

PROBLEM STATEMENT

- Inventory management is a major challenge for T-shirt sellers, as overstocking leads to financial losses, while stockouts result in missed sales.
- Inefficiencies arise due to a lack of data-driven decision-making.
- This project focuses on optimizing inventory by analyzing customer ratings and T-shirt attributes (color, material, fit, design, etc.), helping businesses balance supply and demand effectively.

INDUSTRY APPLICATION

Inventory optimization benefits various fashion and retail sectors:

- E-commerce: Predicts demand using customer ratings.
- Physical Stores: Optimizes shelf space with popular products.
- Print-on-Demand: Refines offerings based on trends.
- Fashion Brands: Streamlines production and distribution.

INDUSTRY OVERVIEW

- Growing Market – The T-shirt industry remains in high demand across various demographics.
- JIT Inventory – Reduces storage costs by restocking based on actual demand.
- Real-Time Tracking – Minimizes errors and improves inventory accuracy.
- Customer Insights – Uses ratings and product features to optimize stock.
- Efficient Supply Chain – Ensures the right products are available at the right time.
- Higher Profits & Satisfaction – Reduces costs, boosts profits, and enhances customer experience.

METHODOLOGY

DATA COLLECTION AND PREPROCESSING

- Customer rating data for T-shirts was manually collected from Amazon. **DATA**
- Features included color, material, fit, design, text elements, and other attributes.
- Each T-shirt was categorized as high or low-rated based on customer reviews.
- The objective was to determine key features influencing high ratings for inventory optimization.

DATA CLEANING AND CONSISTENCY CHECKS

- Handling Missing Values – Removed or imputed missing critical data.
- Standardizing Variables – Ensured uniform naming for categorical features.
- Removing Duplicates – Eliminated redundant entries to avoid bias.
- Outlier Detection – Identified and adjusted extreme values in numerical data.

FEATURE TRANSFORMATION: WEIGHT OF EVIDENCE (WOE)

- Most dataset features were categorical and required numerical conversion.
- WoE Transformation – A technique used to encode categorical variables based on their relationship with the target variable.
- Helps improve model performance by capturing the predictive power of categorical data.
- Ensures better interpretability and handling of categorical features in predictive modeling.

WOE CALCULATION

The WoE transformation is given by:

$$\text{WOE} = \ln\left(\frac{\text{Percentage of High rating instance}}{\text{Percentage of Low rating instance}}\right)$$

WoE helps improve the interpretability of the model and avoids issues associated with one-hot encoding, such as high dimensionality.

STEPWISE REGRESSION FOR FEATURE SELECTION

To ensure only the most relevant features were used in the model, Stepwise Regression was applied. This technique iteratively adds or removes predictor variables based on their statistical significance using Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

THE STEPS INVOLVED:

- Forward Selection
- Backward Elimination
- Hybrid Approach

After applying stepwise regression, only the most statistically significant features were retained for further modeling.

LOGISTIC REGRESSION

Logistic Regression is a statistical method used for binary classification, making it ideal for predicting whether a T-shirt will receive a high or low rating. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities using the sigmoid function:

DERIVATION PART

Consider the sigmoid function,

$$P(Y) = \frac{1}{1+ e^{-(x)}}$$

Applying regression coefficient in sigmod function,

$$P(Y=1)= \frac{1}{1+ e^{-(\beta_0+\beta_1 X_1+ \dots +\beta_k X_k)}} \quad ; \quad P(Y=0)=1-P(Y=1)$$

Consider the odds,

$$\frac{P(Y=1)}{P(Y=0)} \qquad \frac{\frac{1}{1+ e^{-(\beta_0+\beta_1 X_1+ \dots +\beta_k X_k)}}}{1 - \frac{1}{1+ e^{-(\beta_0+\beta_1 X_1+ \dots +\beta_k X_k)}}}$$

$$\frac{P(Y=1)}{P(Y=0)} = e^{(\beta_0+\beta_1 X_1+ \dots +\beta_k X_k)}$$

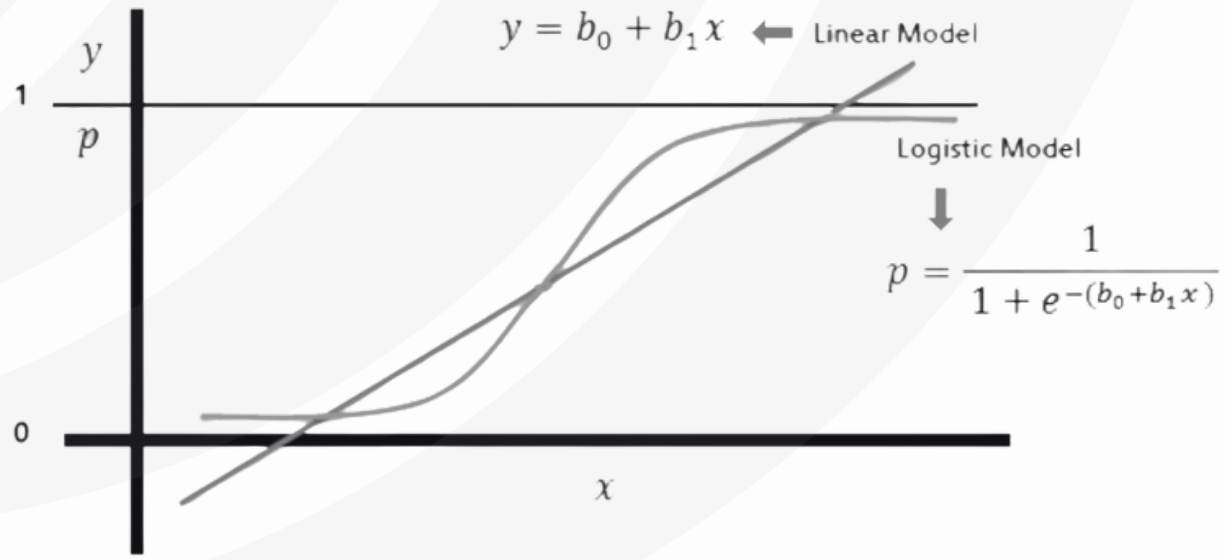
Where P(Y=1) = p and P(Y=0) = 1-P

$$\text{Log}\left(\frac{P}{1-P}\right) = \beta_0+\beta_1 X_1+ \dots +\beta_k X_k$$

- P(Y=1|X) is the probability of a T shirt receiving a high rating.
- β_0 is the intercept
- $\beta_1,\beta_2,...\beta_k$ are the coefficient for the predictor variable.

DERIVATION OF LOGISTIC REGRESSION COEFFICIENTS

The coefficients (β) of logistic regression are estimated using Maximum Likelihood Estimation (MLE). Given a dataset with non observations and predictors X, the likelihood function is:



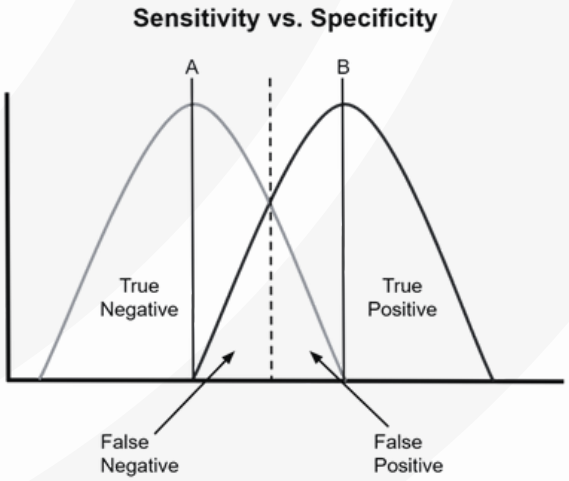
MODEL EVALUATION METRICS

Once the logistic regression model was built, its performance was evaluated using various metrics

Confusion Matrix:-

Actual / Predicted	High Rating (1)	Low Rating (0)
High Rating (1)	True Positive (TP)	False Negative (FN)
Low Rating (0)	False Positive (FP)	True Negative (TN)

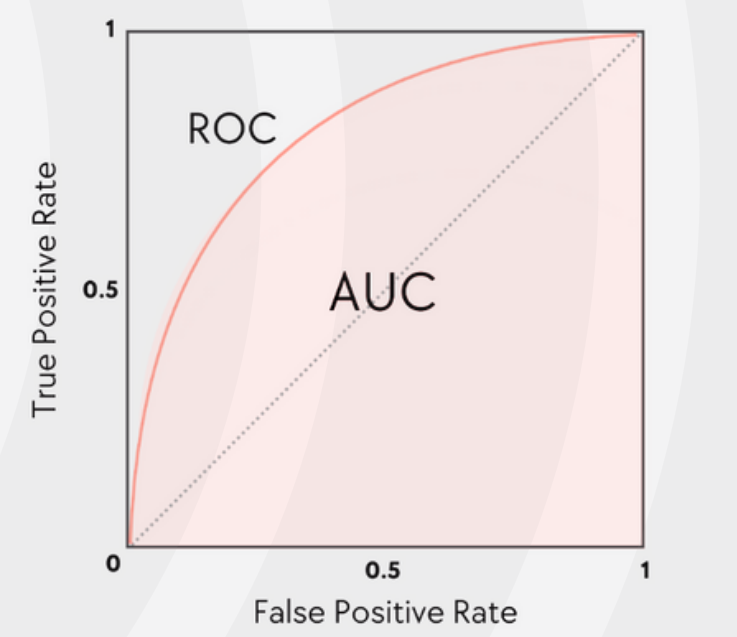
Sensitivity & Specificity:-



- Sensitivity measures how well the model correctly identifies high-rated T-shirts.
- Specificity measures how well the model correctly identifies low-rated T-shirts

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \quad ; \quad \textbf{Specificity} = \frac{TN}{TN+FP}$$

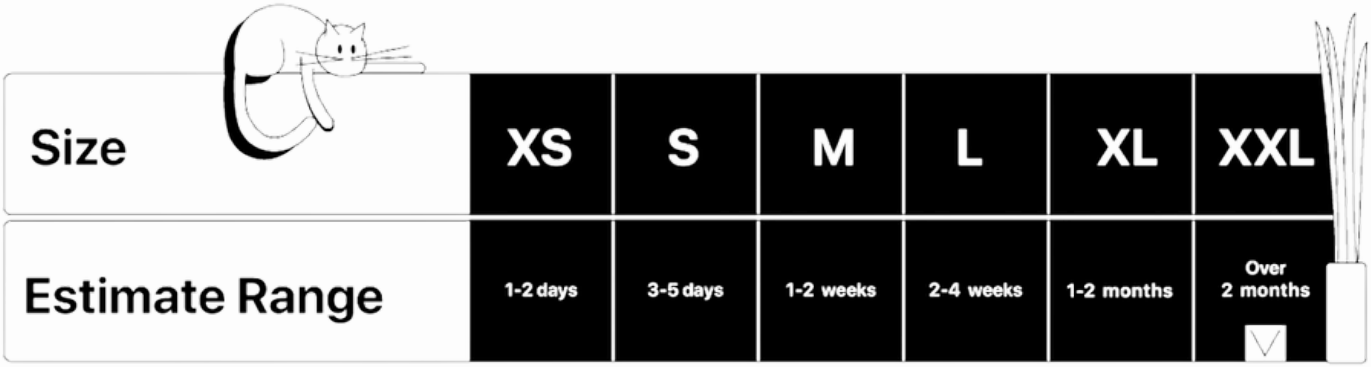
ROC Curve and AUC:-



The ROC Curve plots Sensitivity vs. 1 - Specificity, showing the model's ability to distinguish between high and low ratings.

The Area Under the Curve (AUC) quantifies model performance:

- AUC = 1 → Perfect classifier
- AUC = 0.5 → Random guess
- Higher AUC values indicate better performance



Decision Tree:-

A Decision Tree is a classification model that splits data into branches based on feature values. Each internal node represents a decision rule, each branch represents an outcome, and each leaf node represents a final classification.

- **Decision Tree Splitting Criteria**

To determine the best splits, the following metrics were used:
Gini Index
Measures impurity in a node:

$$Gini = 1 - \sum_{i=1}^{C} P_i^2$$

Measures the reduction in uncertainty after a split:

- **Entropy (Information Gain)**

$$Entropy = -\sum_{i=1}^{C} P_i \log_2 P_i$$

Where P is the probability of class . A lower Gini or Entropy value indicates a better split.

- **Decision Tree Visualization**

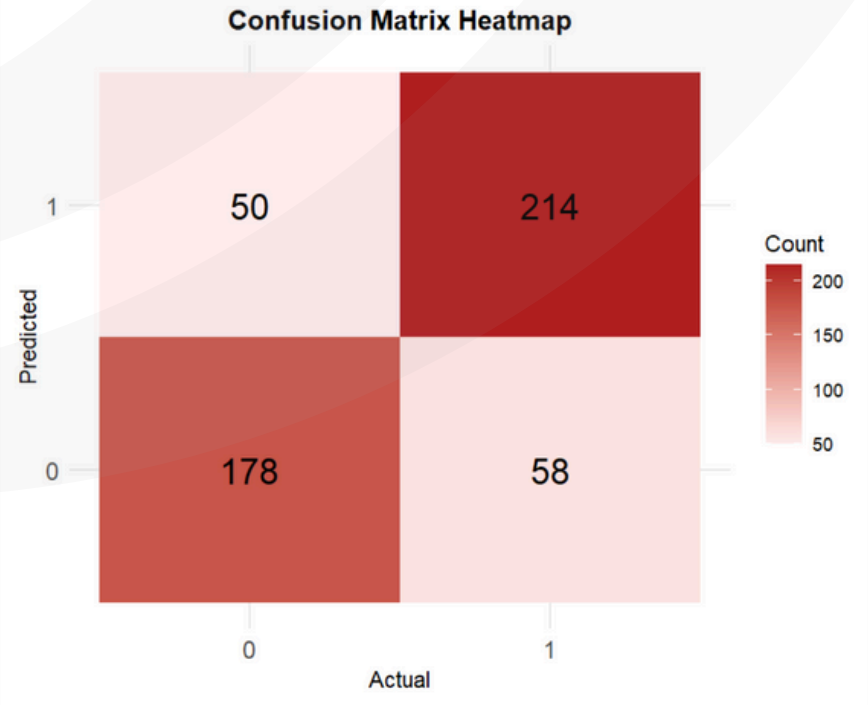
A decision tree visualization was created to understand how different T-shirt features influence customer ratings. This provides insights into which factors contribute most to high ratings and low ratings.

ANALYSIS

BINARY LOGISTIC REGRESSION

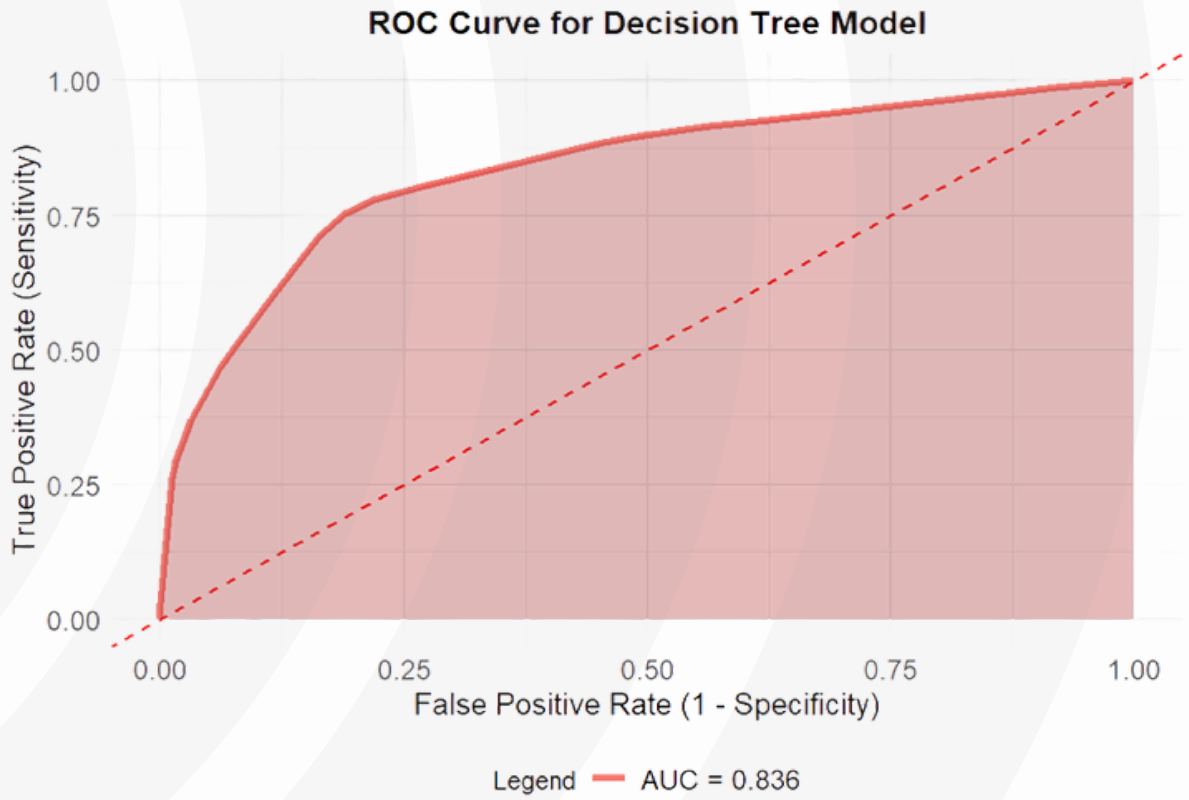
VARIABLE	Estimate	Std.Error	Z Value	Pr (> |z|)	Odds Ratio	CI Lower bound	CI Upper bound
(Intercept)	-1.0719461	0.2875056	-3.728	0.000193	0.342	0.192	0.594
Price	0.0024337	0.0005587	4.356	1.32E-05	1.002	1.001	1.004
major_colour	-0.9478848	0.2417339	-3.921	8.81E-05	0.388	0.240	0.620
brand_name	-1.5195482	0.4742031	-3.204	0.001353	0.219	0.085	0.550
sleeve_type	-2.4233452	0.6668140	-3.634	0.000279	0.089	0.023	0.322
design_location	-0.7311179	0.3214946	-2.274	0.022959	0.481	0.252	0.895
main_colour	-1.2160816	0.4542629	-2.677	0.007427	0.296	0.119	0.712
Multicolour	-1.3434721	0.6400491	-2.099	0.035816	0.261	0.071	0.894
fit_type	-0.7640160	0.2960984	-2.580	0.009872	0.466	0.259	0.830
text	4.5786376	1.3712643	3.339	0.000841	97.382	7.155	1580.224
text_location	-0.9871084	0.4197697	-2.352	0.018695	0.373	0.150	0.803
message_grouping	-0.8670900	0.4106940	-2.111	0.034748	0.420	0.184	0.928

CONFUSION MATRIX

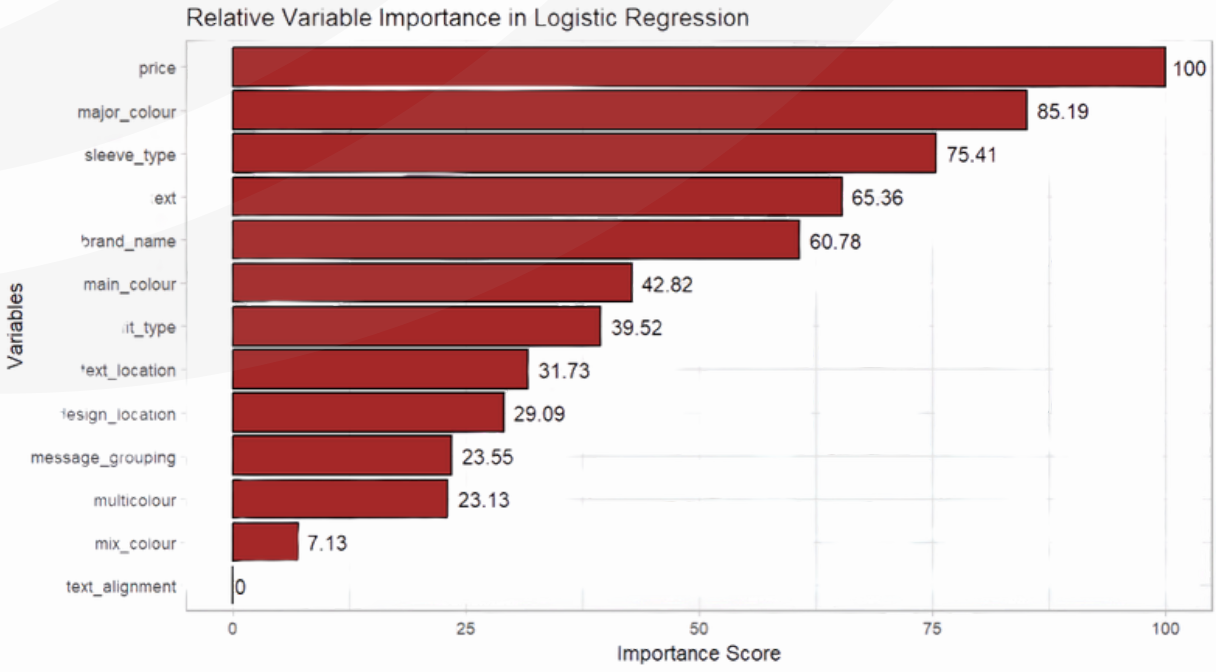


- **Sensitivity :** 0.754
- **Specificity :** 0.811
- **Accuracy** : 0.784

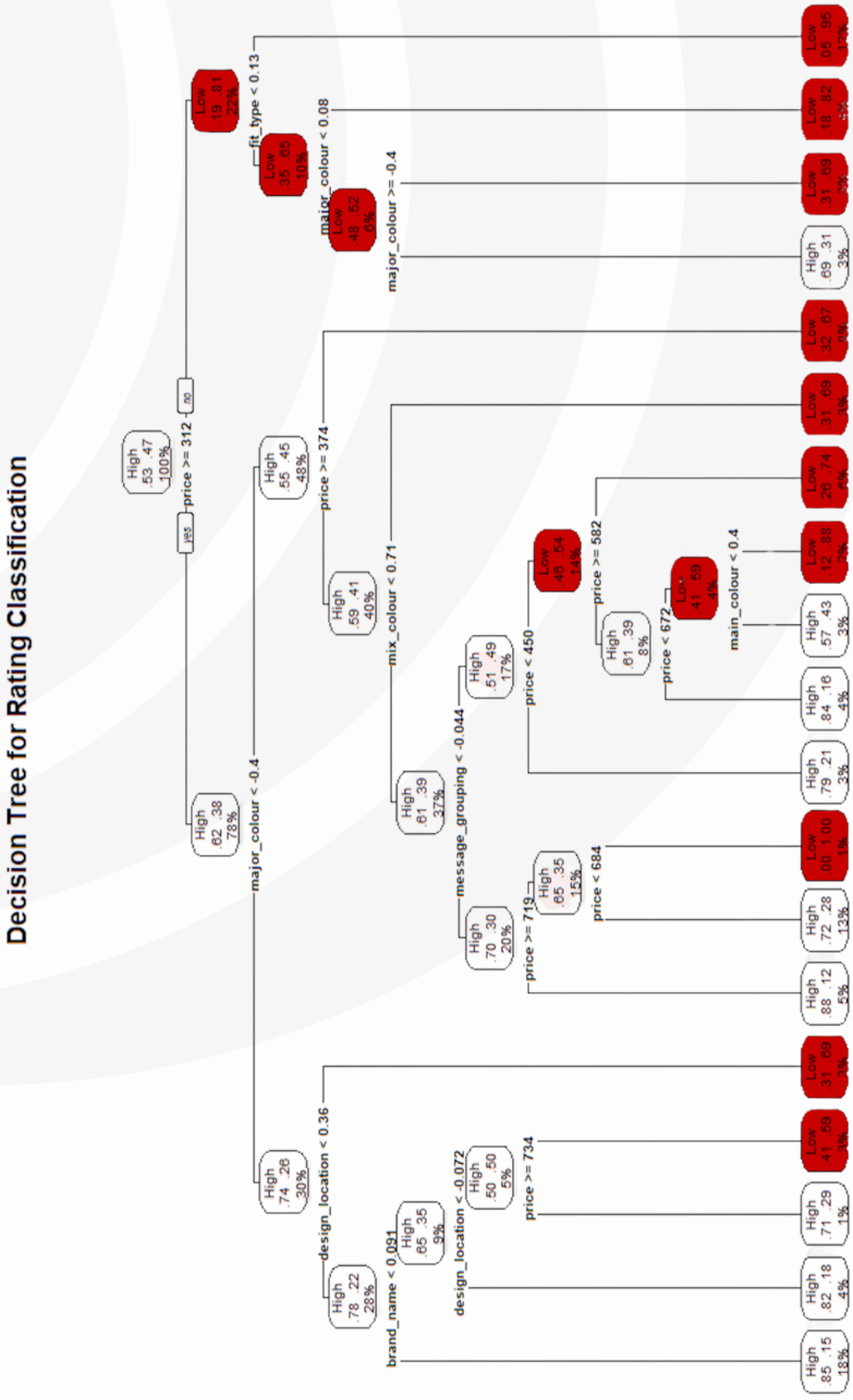
ROC CURVE



RELATIVE VARIABLE IMPORTANCE



DECISION TREE



TOP TWO SEGMENTS FOR HIGH & LOW PREFERENCE

- Based on recent findings, we've identified key customer preferences when it comes to T-shirt.
- Understanding these can help stock wisely and enhance customer satisfaction. Here are the two segments to focus on high demand items.

PRICING & COLOURS

Customer prefer T-shirts priced at 312 or above ,associating high prices with better quality. Stocking black and blue T-shirt in this range will align with their preference .

DESIGN PLACEMENT INSIGHTS

Design placement is crucial; customer favour T-shirt with design on the black, Sleeves, or none at all. Avoid heavily designed fronts unless there's clear demand.

COLOURS TEXT OPTIONS

Opt for subtle text colours like grey, redder yellow, as customer prefer these over vibrant hues. Focus on T-shirts with neutral test to see their taste .

FOCUS ON FIT TYPES

Regular fit T-shirt are less favoured; stock alternative fits like slim or oversized together meet customer expectations and enhance satisfaction.

THANK YOU!

suryavincent19@gmali.com 8124705377

naveenkumarsenthilkumar01@gmail.com 8838491422