Analyzing and Classifying Customer Returns with Machine Learning Models

Group 6: Muhammet Ali Buyuknacar, Zhiwei Gu, Jialong Li, Siyan Li

Introduction

Client: Dillard's
American department store chain

Business question: **Predict product returns**
based on product information and transaction record

Goal: Optimize inventory management strategies
Maximize Return on Investment

Data Ingestion

- Migrated the dataset into cloud SQL database and set the constrains
- Dropped the faulty rows with extra columns
- Figured out the sequence of columns
- Gathered external data regarding social economic factors from government official website

EDA

- Transaction data across 2 years, 120 million records
- 15560300 SKUs
- 31 states, 299 cities, 391 zipcodes, 453 stores

- Dropped rows with missing Cost and Retail
- Matched the Purchases with Returns
- 19,101,307 purchases, 1,011,886 returns, imbalanced dataset

Feature Engineering

Color, Size

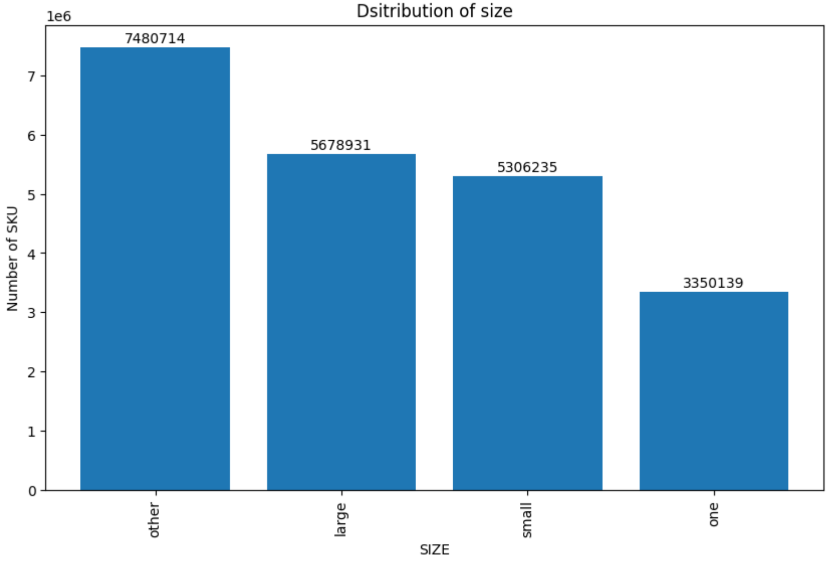
- Mapping
- Color: 71322 color --> 13 groups
- Size: 9784 size --> 4 groups

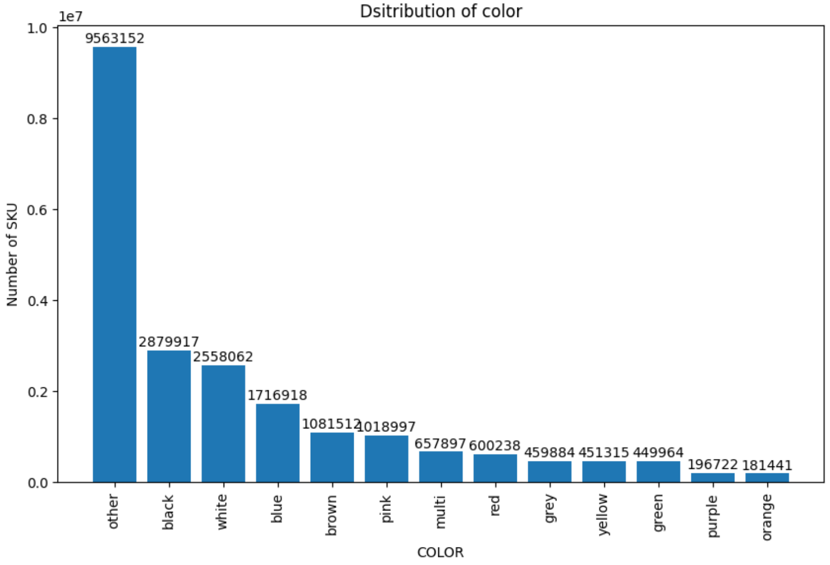
Date, Price

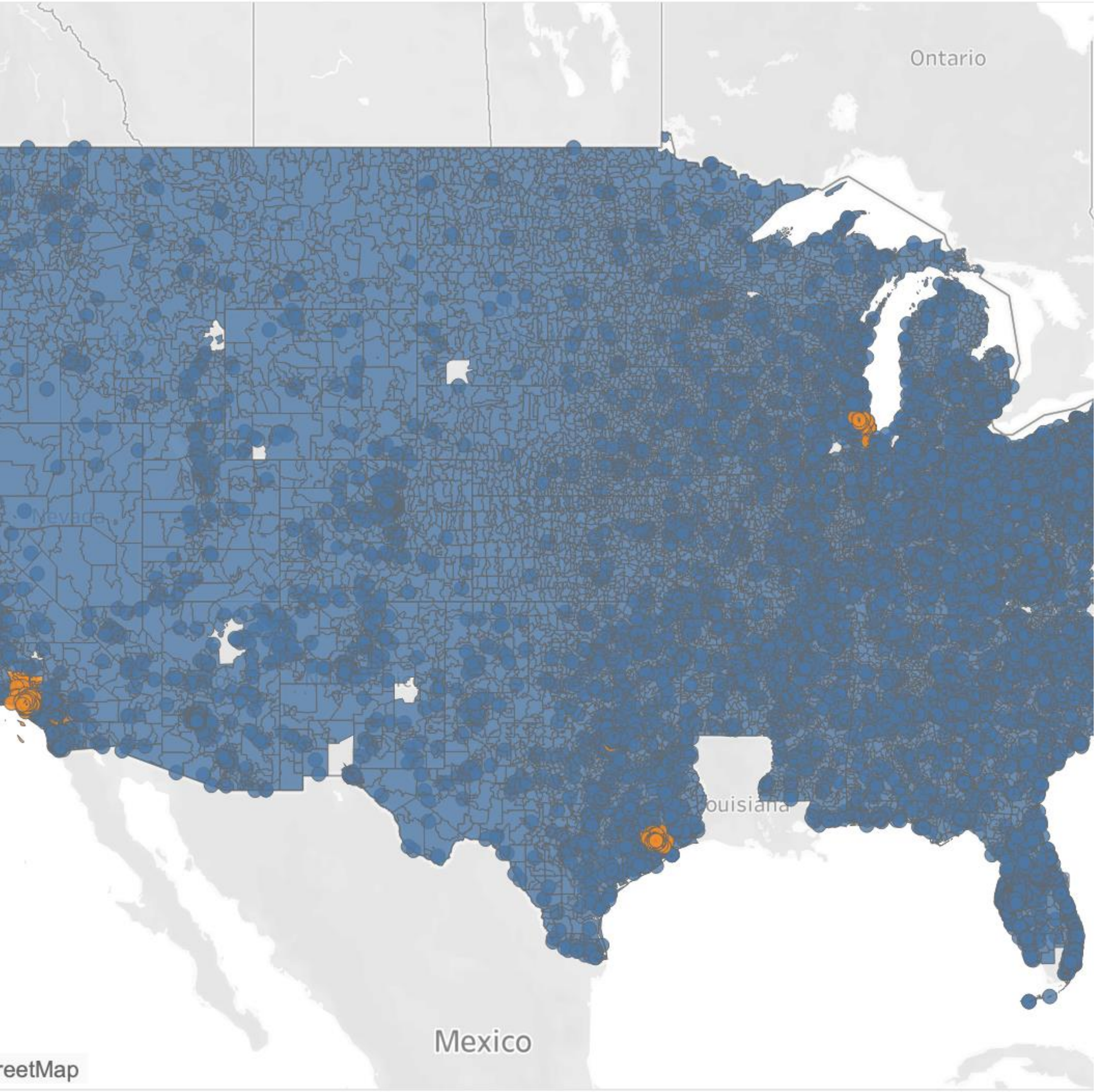
- Holiday or not, Weekend or not
- Discount or not

Socioeconomic factor

- GDP, population, poverty rate, median income
- Clustering



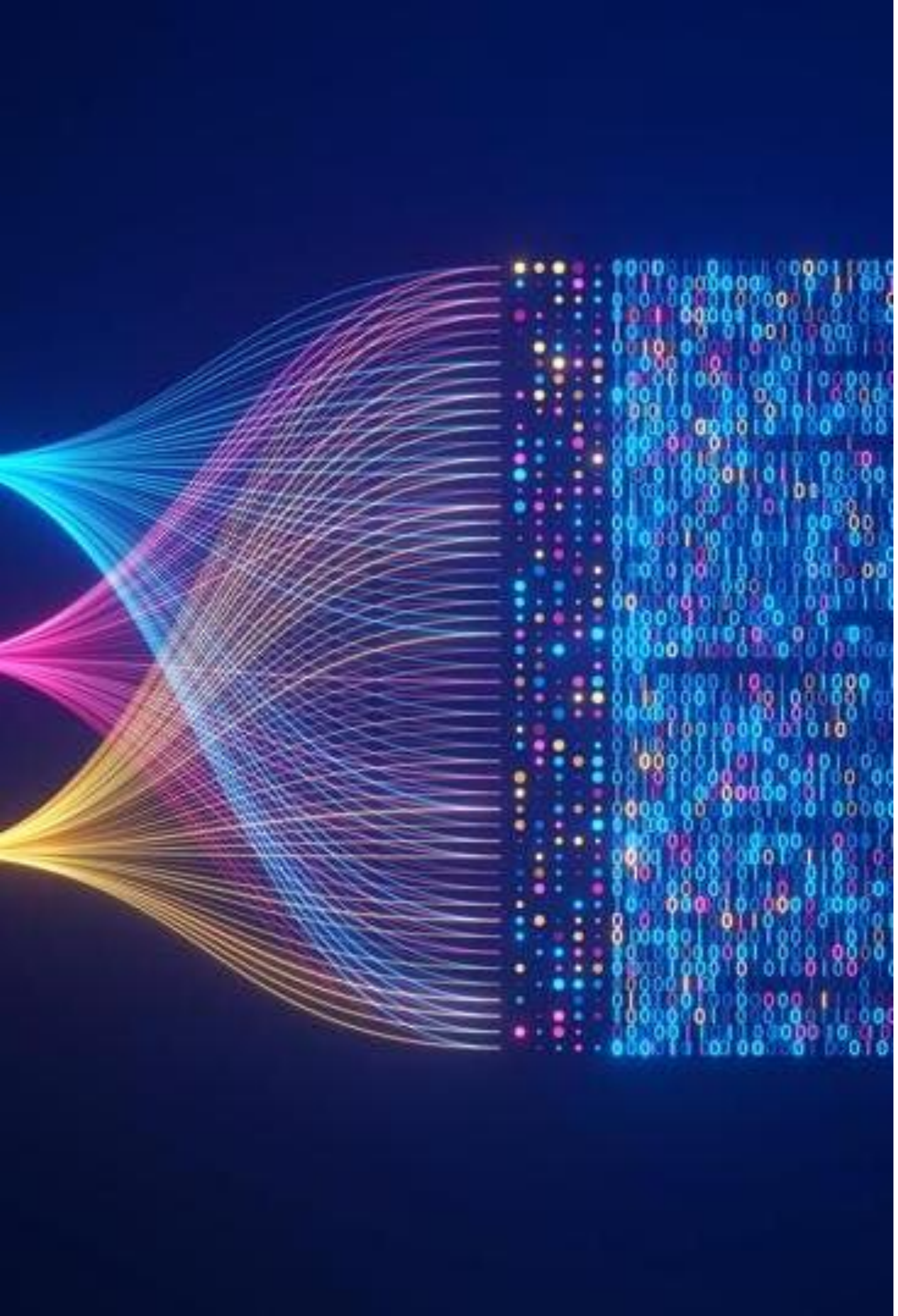




(generated) and Latitude (generated). Color shows details about Econ Label. Details are shov

Feature Engineering

- Social-economic factors
- Use zip code as the smallest granularity of location representation
- Some zip codes represents multiple counties, aggregate within the same zip code
- Use GDP, poverty, population, and median income to cluster the zip codes into two groups
- 36483 in group 0, 1092 in group 1

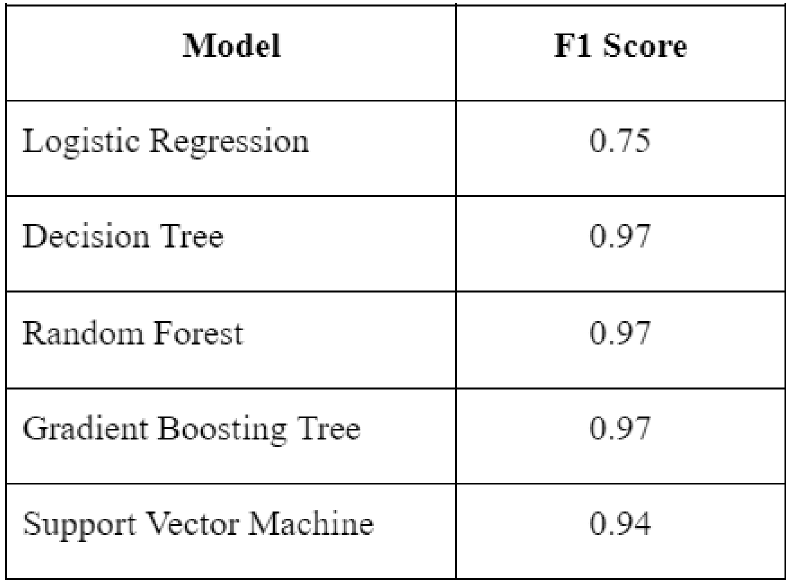


Modeling

- **Logistic Regression**
  - Employed binomial logistic regression
  - Almost all features are statistically significant
- **Tree-Based Algorithms**
  - Employed decision tree, random forest, and Gradient Boosting Tree
  - Compared their performances on the validation set with different depths
- **Support Vector Machine**
  - SVM is good at classification tasks
  - We use validation and find out kernel 'rbf' has the best performance

Model Benchmarking

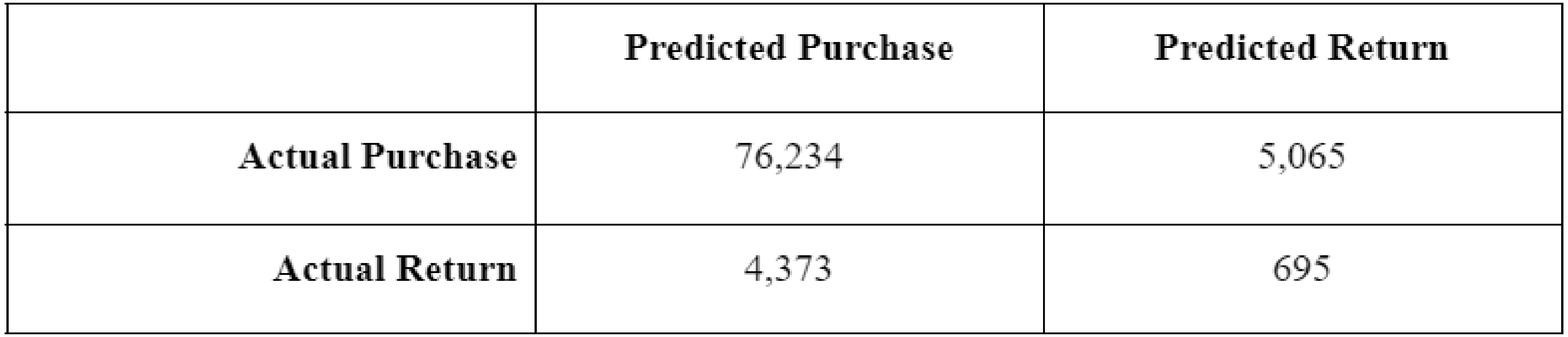
- Performances by F1 Score



Model	F1 Score
Logistic Regression	0.75
Decision Tree	0.97
Random Forest	0.97
Gradient Boosting Tree	0.97
Support Vector Machine	0.94

Model Benchmarking

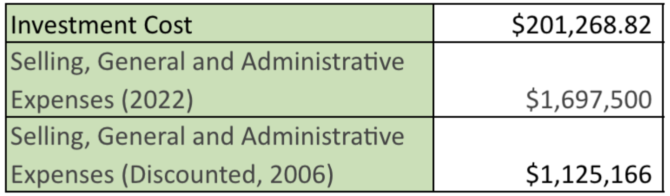
- We choose SVM with the "rbf" kernel as the final strategy, given that it has a more balanced prediction
- This is the confusion matrix:



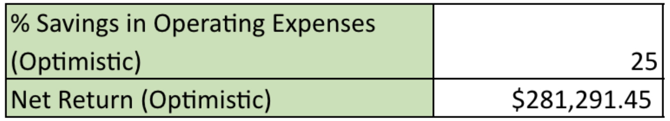
	Predicted Purchase	Predicted Return
Actual Purchase	76,234	5,065
Actual Return	4,373	695

ROI Analysis

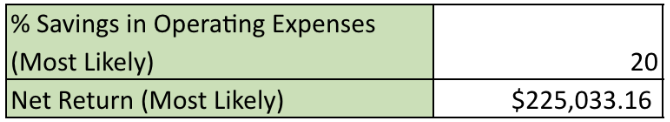
- Calculated investment costs by salary and computing costs
- Simulated 3 different scenarios
- Calculated the expected savings from operating expenses
- The expected ROI is 11.81%



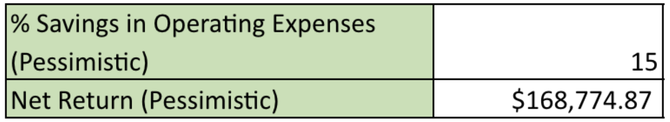
Investment Cost	$201,268.82
Selling, General and Administrative Expenses (2022)	$1,697,500
Selling, General and Administrative Expenses (Discounted, 2006)	$1,125,166



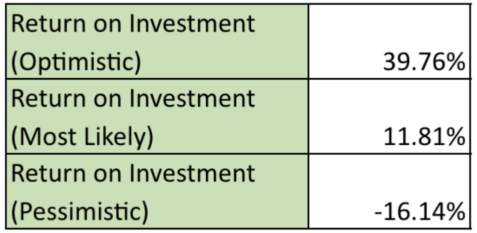
% Savings in Operating Expenses (Optimistic)	25
Net Return (Optimistic)	$281,291.45



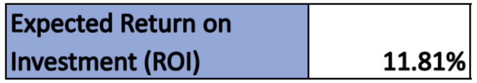
% Savings in Operating Expenses (Most Likely)	20
Net Return (Most Likely)	$225,033.16



% Savings in Operating Expenses (Pessimistic)	15
Net Return (Pessimistic)	$168,774.87

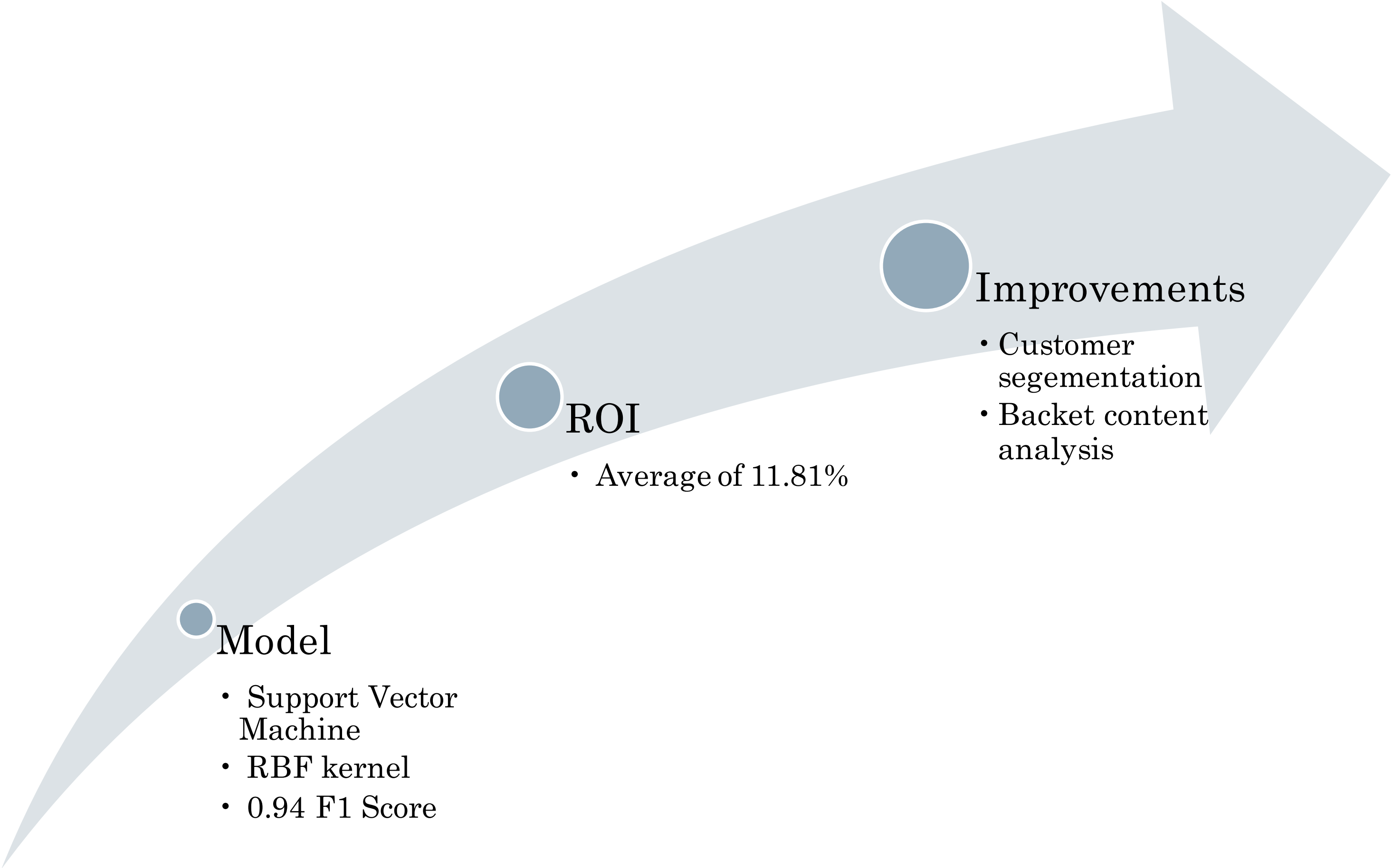


Return on Investment (Optimistic)	39.76%
Return on Investment (Most Likely)	11.81%
Return on Investment (Pessimistic)	-16.14%



Expected Return on Investment (ROI)	**11.81%**

Conclusion



Model

- Support Vector Machine
- RBF kernel
- 0.94 F1 Score

ROI

- Average of 11.81%

Improvements

- Customer segementation
- Backet content analysis