Walmart Customer Trip Type Classification

Carnegie Mellon University
Information Systems Management
Business Intelligence and Data Analytics

95-828 Machine Learning for Problem Solving

Instructor: Leman Akoglu
Team member: Ya Ting Chang; Michelle Hsu

Image: https://commons.wikimedia.org/wiki/File:Walmart_logo.svg

Motivation

Customers come to Walmart for various purposes

Different marketing strategy is applied to different types of customer visit

Images:
(1)shopping by Mohamad Arif Prasetyo from the Noun Project
(2) https://es.digitaltrends.com/celular/devoluciones-walmart-aplicacion/

Problem Definition

Problems: Walmart’s customer trip type is categorized into **38** types

→ Initiate effective marketing strategy

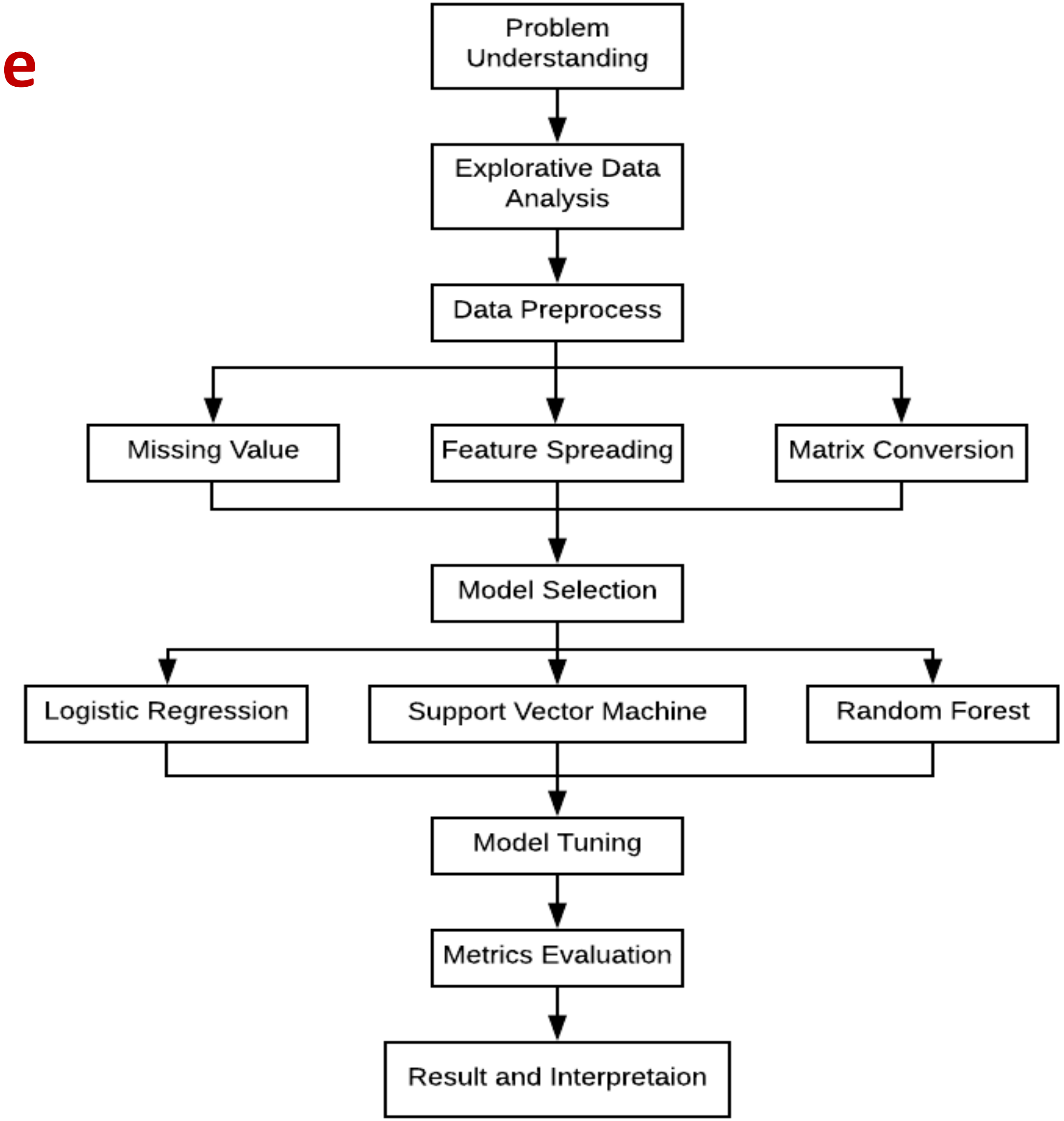
→ Boost sales and reduce marketing cost

Objective: Predict customer trip type with higher accuracy

Supervised Learning → Multiclass Classification

Image:
(1)https://www.iconfinder.com/icons/3018530/buyer_customer_figure_person_portrait_profile_user_icon#size=128
(2)https://www.iconfinder.com/icons/49832/help_question_mark_icon#size=130

Project Pipeline



Problem Understanding

Explorative Data Analysis

Data Preprocess

Missing Value

Feature Spreading

Matrix Conversion

Model Selection

Logistic Regression

Support Vector Machine

Random Forest

Model Tuning

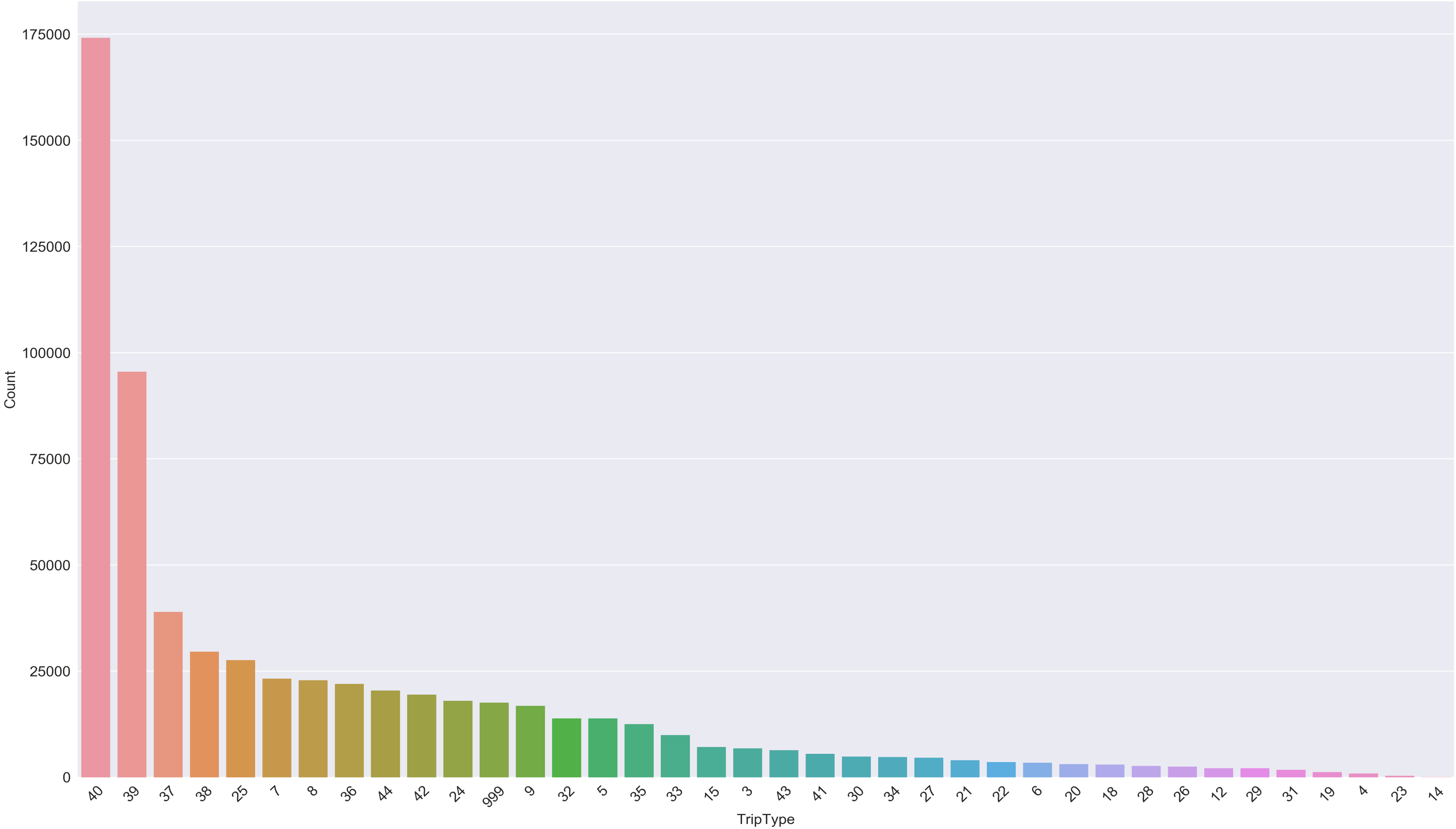
Metrics Evaluation

Result and Interpretaion

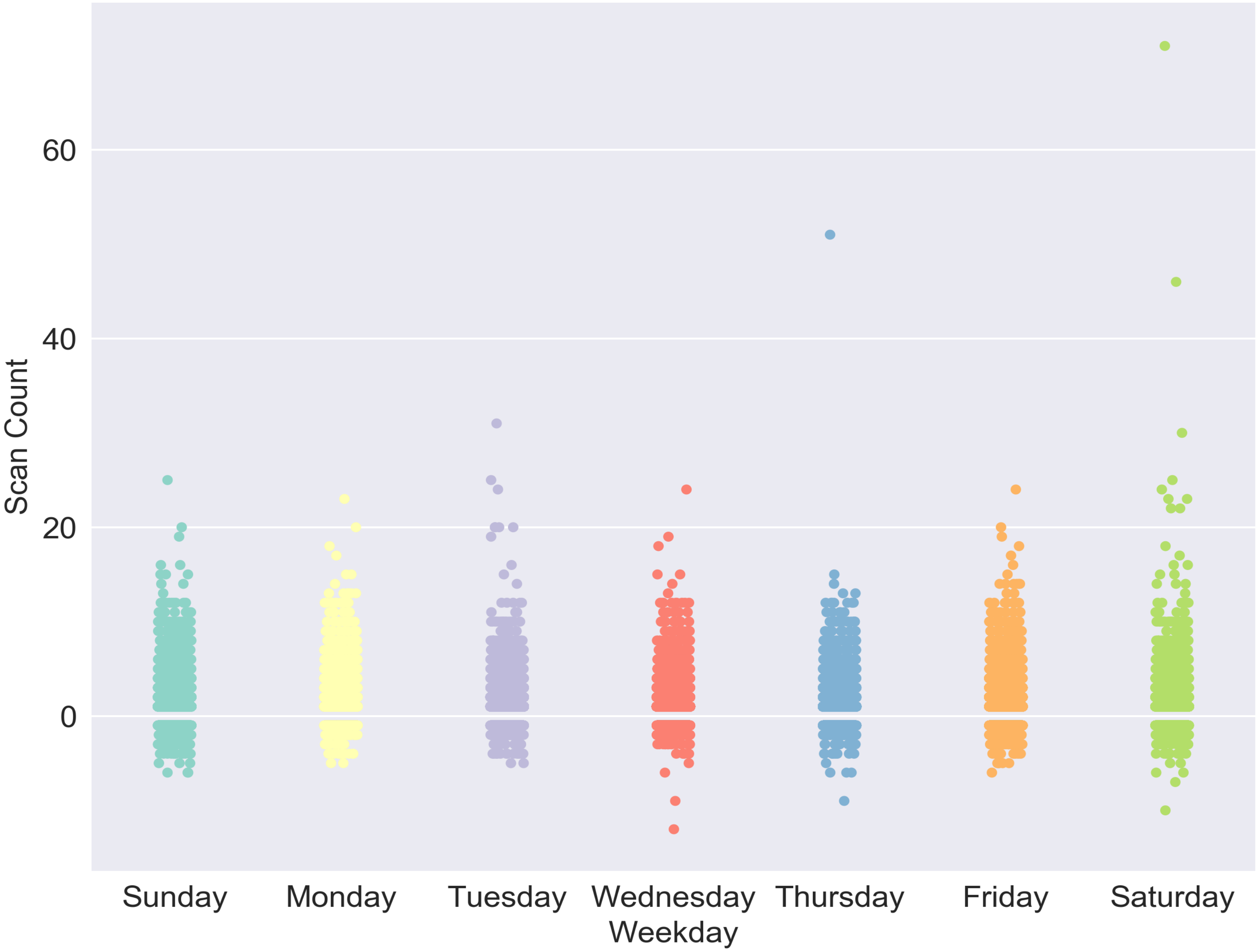
Features

Feature	Description
Department Description	A high-level description of the product’s department
FinelineNumber	A refined category for the product purchased
ScanCount	The number of the given item that was purchased A negative value indicates a product return
TripType	The 38 original trip types. TripType_999 is an "other" category
VisitNumber	An id corresponding to a single trip by a single customer
Weekday	The weekday of the customer visit

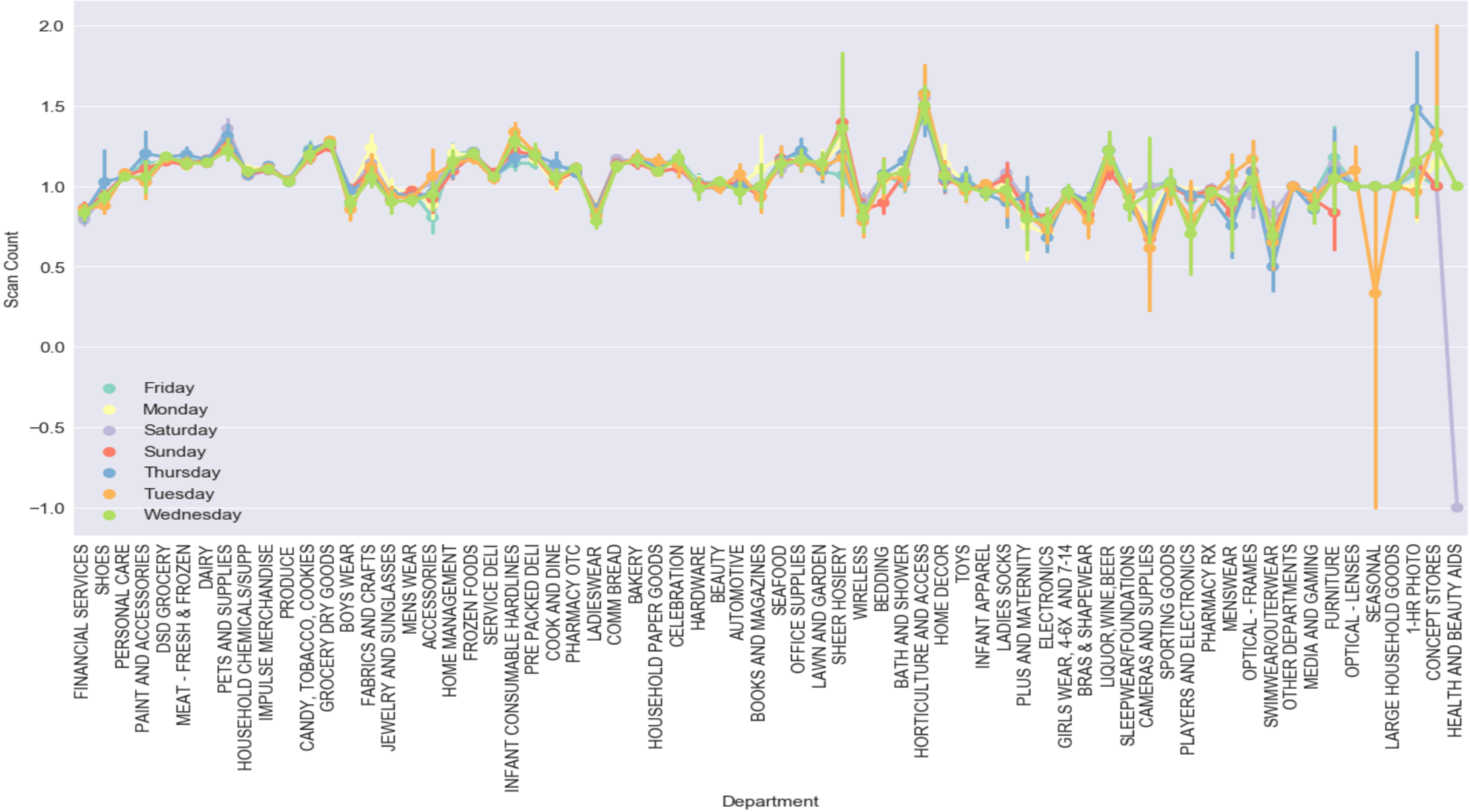
Explorative Data Analysis



Explorative Data Analysis

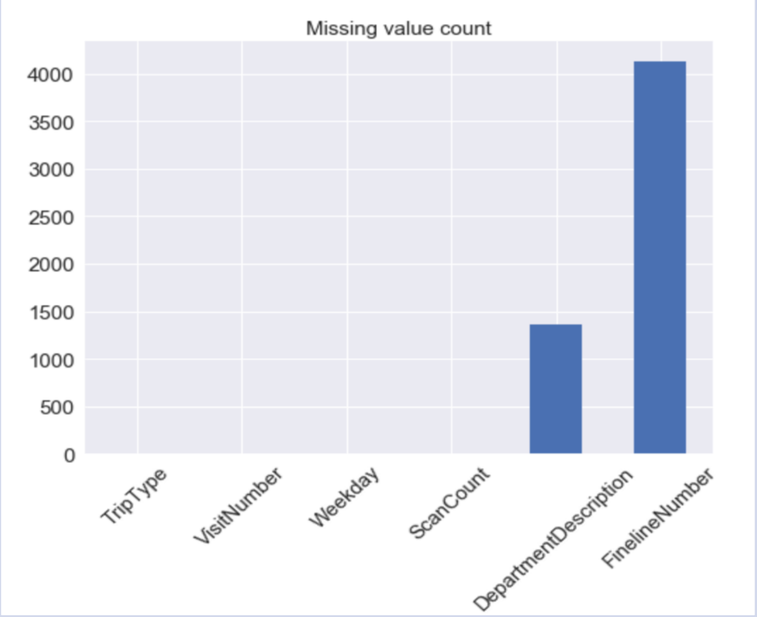


Explorative Data Analysis



Data Preprocess

Missing Value



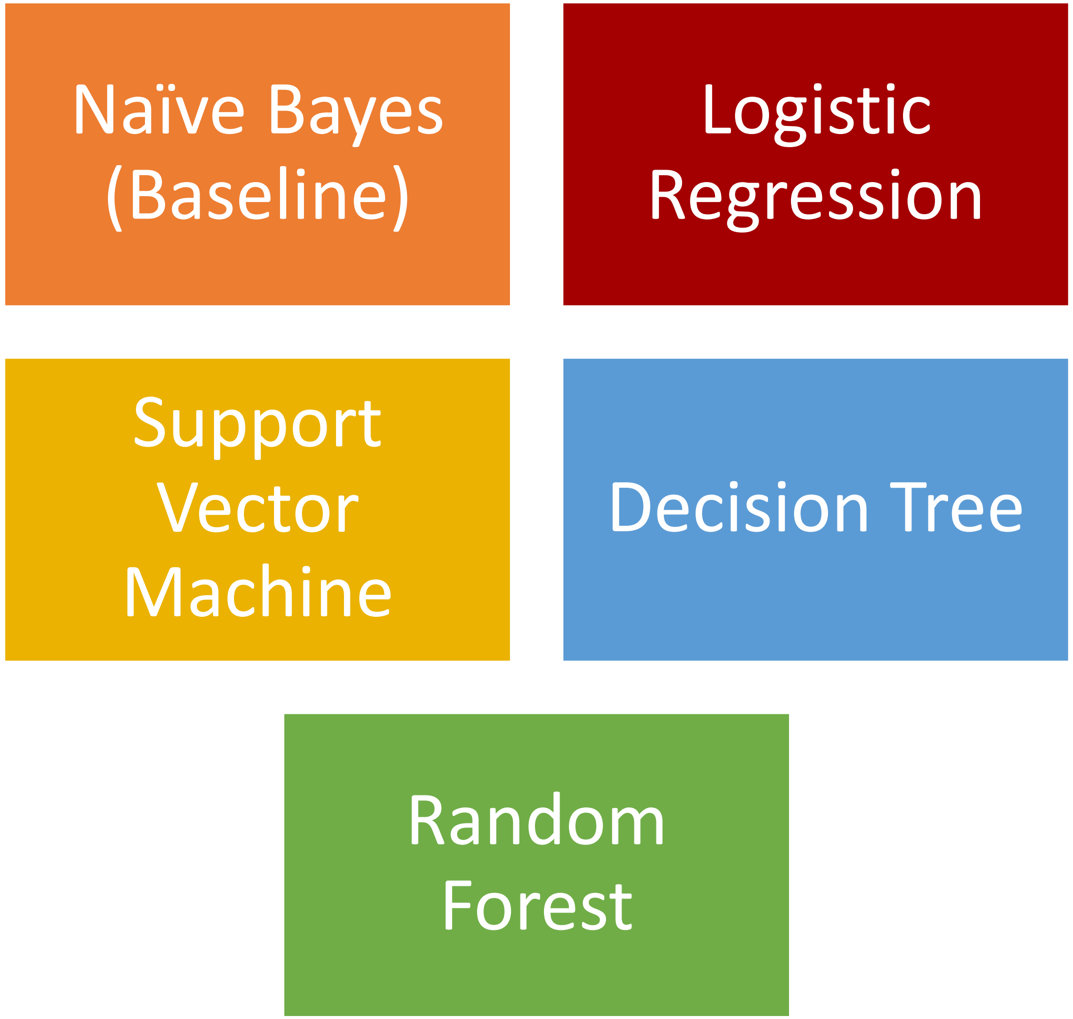
Feature Spreading

- FinelineNumber
- Weekday
- DepartmentDescription
- ScanCount: return, few, medium, many

Matrix Transformation

- Dense to Sparse
- 38,206 rows
- 5275 columns

Preliminary Model Selection



Naïve Bayes (Baseline)

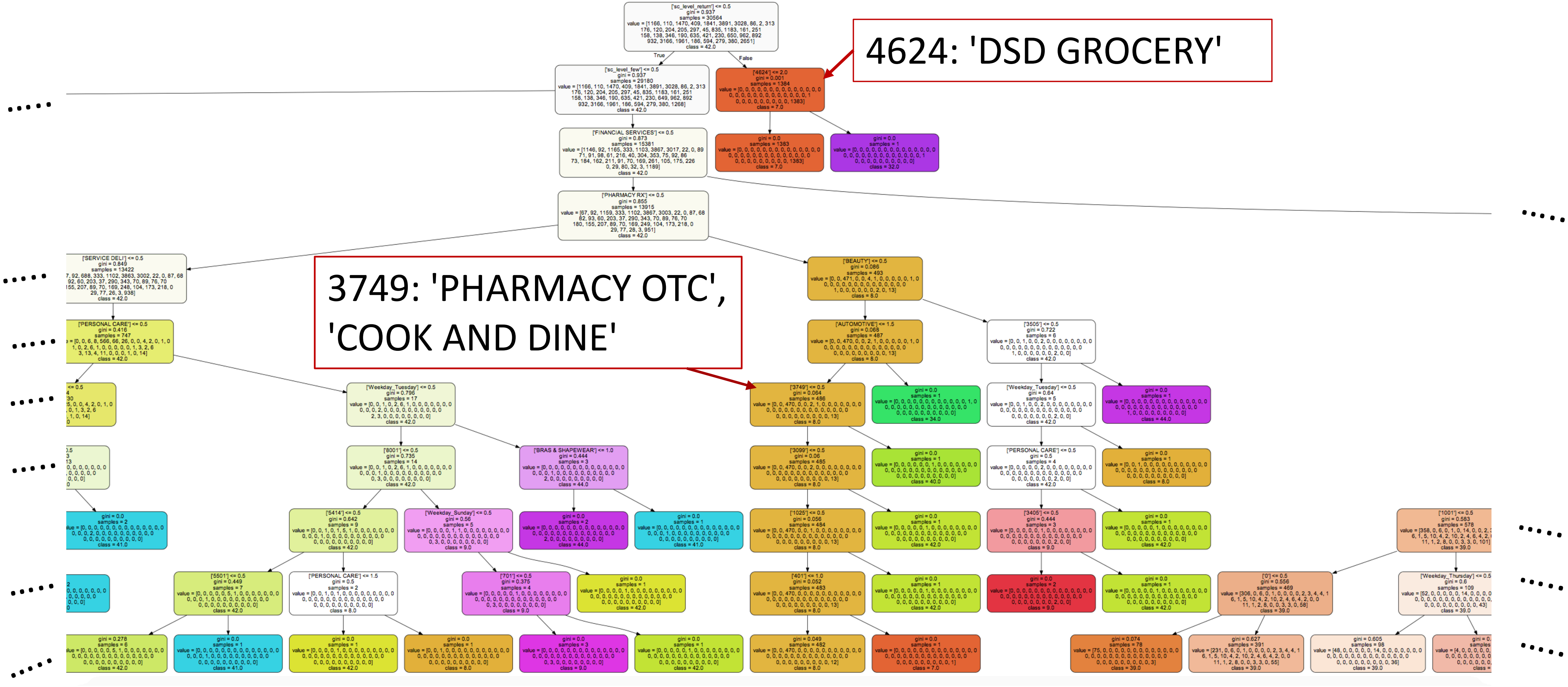
Logistic Regression

Support Vector Machine

Decision Tree

Random Forest

Decision Tree plot



4624: 'DSD GROCERY'

3749: 'PHARMACY OTC', 'COOK AND DINE'

Model Tuning

5-Fold Cross Validation

Logistic Regression

- C
- regularization strength

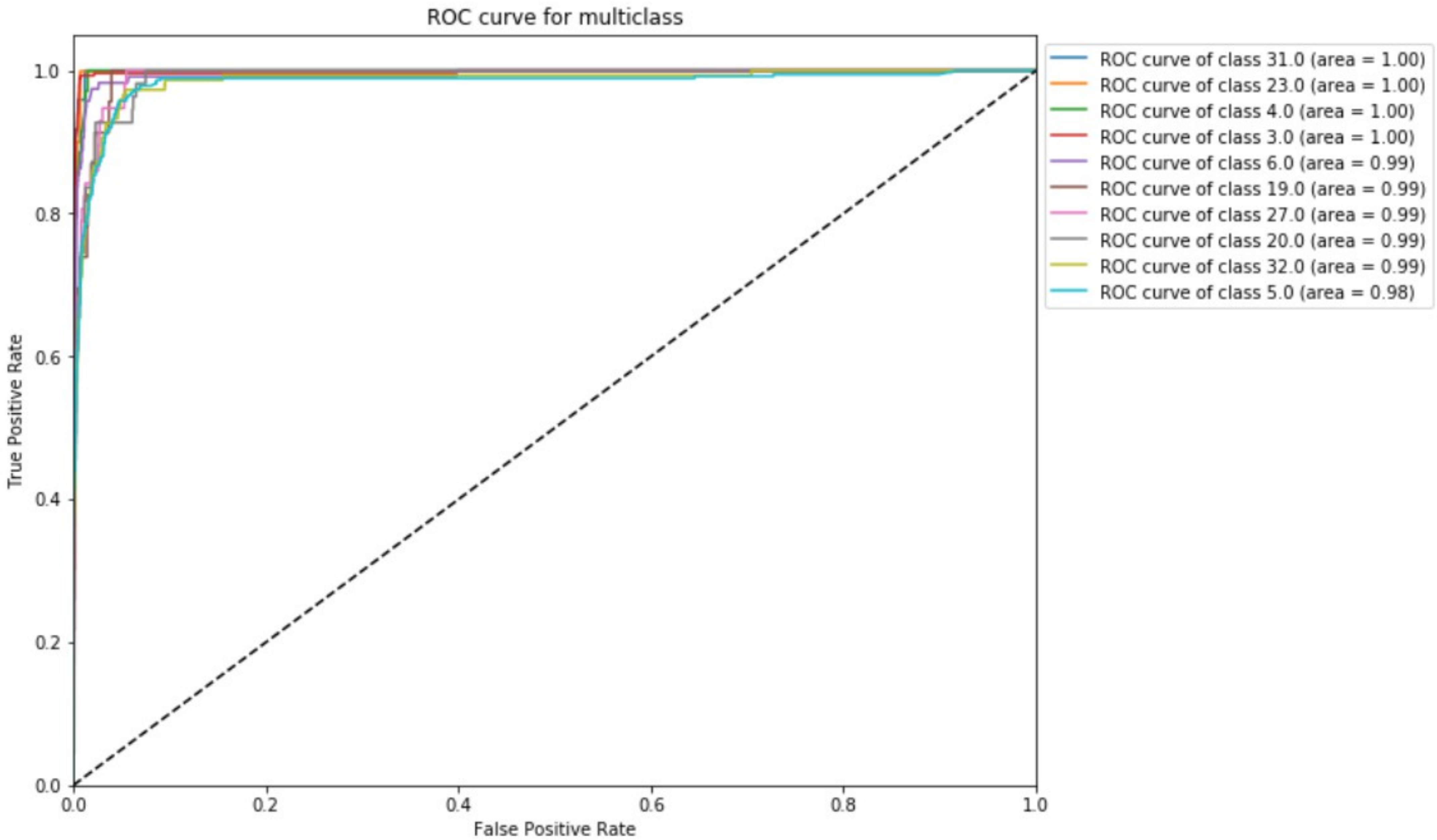
Support Vector Machine

- C
- Penalty of the error term

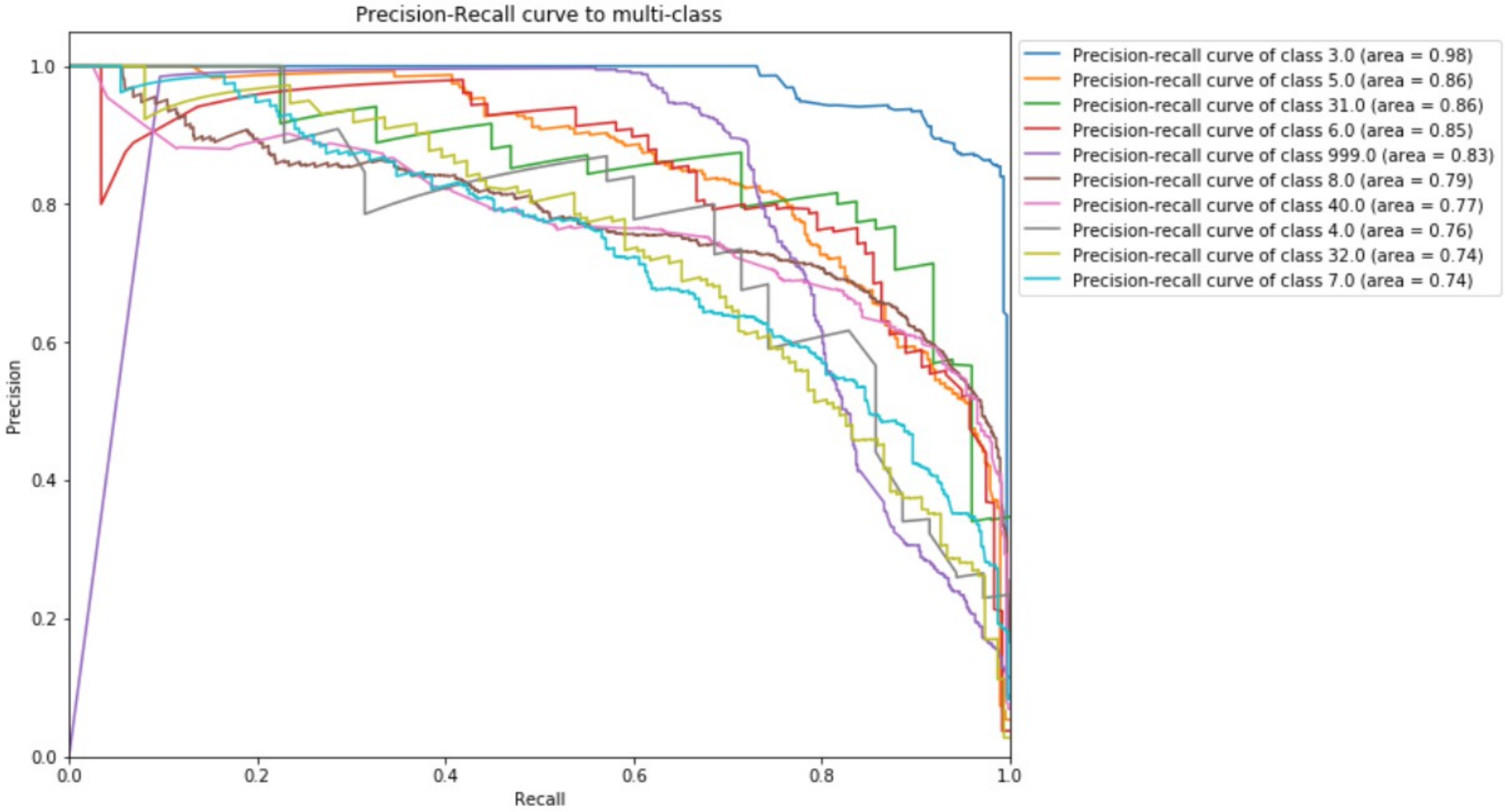
Random Forest

- #trees
- #max_depth

Model Evaluation - Random Forest (ROC curve)



Model Evaluation - Random Forest (Precision-Recall Curve)



Model Comparison

	Accuracy	F1-Score
Naïve Bayes	0.6154	0.6010
Logistic Regression	0.6710	0.6589
SVM	0.6705	0.6566
Random Forest	0.6110	0.5820

Limitations & Future Work

Limitations:

1. Cannot interpret the underlying meaning of FinelineNumber and TripType
2. Imbalanced data for multiclass classification is hard to handle. SMOTE is not robust to the multiclass setting unless converted to binary classification task

Future Work:

1. Use the whole dataset, which is comprised of 95,828 instances and 5,275 columns, to train the model
2. Try to tune other hyperparameters or methods to see the whether the performance can be boosted