**DATASCI 450**

**Class Project | Group 10 | Part 1 & 2**

# **Project Title**

***Walmart Recruiting: Trip Type Classification***

# **Team Member**

***Aleksey Kramer, Jagger Bodas, Winston Featherly-Bean***

**TABLE OF Contents**

[**Project Title** 1](#_Toc436646947)

[**Team Member** 1](#_Toc436646948)

[**Project Description** 3](#_Toc436646949)

[**Objective and Scope** 3](#_Toc436646950)

[**PROJECT: Part 1** 4](#_Toc436646951)

[**Exploratory Data Analysis and Data Issue Identification (Questions a and b)** 4](#_Toc436646952)

[**TOTAL NUMBER OF RECORDS** 4](#_Toc436646953)

[**SUMMARY STATISTICS FOR EACH COLUMN** 4](#_Toc436646954)

[**OBSERVATIONS** 7](#_Toc436646955)

[**HIGH LEVEL ANALYTIC PROBLEM (Question C)** 7](#_Toc436646956)

[**PROJECT: Part 2** 8](#_Toc436646957)

[**PROJECT: Part 3** 8](#_Toc436646958)

[**REFERENCES / ATTACHMENTS** 8](#_Toc436646959)

[**PROJECT Part 1 Attachments** 8](#_Toc436646960)

[**Project Part 2 Attachments** 9](#_Toc436646961)

# **Project Description**

Walmart uses both art and science to continually make progress on their core mission of better understanding and serving their customers. One way Walmart is able to improve customers' shopping experiences is by segmenting their store visits into different trip types.

Whether they're on a last minute run for new puppy supplies or leisurely making their way through a weekly grocery list, classifying trip types enables Walmart to create the best shopping experience for every customer.

Currently, Walmart's trip types are created from a combination of existing customer insights ("art") and purchase history data ("science"). In their third recruiting competition, Walmart is challenging Kagglers to focus on the (data) science and classify customer trips using only a transactional dataset of the items they've purchased. Improving the science behind trip type classification will help Walmart refine their segmentation process.

# **Objective and Scope**

For this competition, we are tasked with categorizing shopping trip types based on the items that customers purchased. To give a few hypothetical examples of trip types: a customer may make a small daily dinner trip, a weekly large grocery trip, a trip to buy gifts for an upcoming holiday, or a seasonal trip to buy clothes.

Walmart has categorized the trips contained in this data into 38 distinct types using a proprietary method applied to an extended set of data. We are challenged to recreate this categorization/clustering with a more limited set of features. This could provide new and more robust ways to categorize trips.

The training set (train.csv) contains a large number of customer visits with the TripType included. You must predict the TripType for each customer visit in the test set (test.csv). Each visit may only have one TripType. You will not be provided with more information than what is given in the data (e.g. what the TripTypes represent or more product information).

# **PROJECT: Part 1**

Define the objective and scope of the project. Gather and organize data for the project.

1. Conduct exploratory data analysis such as visualizing the data through graphs, tables, summary statistics, and other means to understand the data.
2. Identify any issues associated with data gap, data size, data type, data manipulation, data storage, and data retrieval for analysis. Structured or unstructured data?
3. Describe the high level analytic problem needs to be resolved: supervised learning, unsupervised learning.

## **Exploratory Data Analysis and Data Issue Identification (Questions a and b)**

### **TOTAL NUMBER OF RECORDS**



### **SUMMARY STATISTICS FOR EACH COLUMN**

|  |
| --- |
| **TripType** |
|  |

|  |  |
| --- | --- |
| **VisitNumber** | |
|  |  |

|  |  |
| --- | --- |
| **WeekDay** | |
|  |  |

|  |
| --- |
| **UPC** |
|  |

|  |
| --- |
| **ScanCount** |
|  |

|  |  |
| --- | --- |
| **Department Description** | |
|  |  |

|  |
| --- |
| **FileLineNumber** |
|  |

### **OBSERVATIONS**

We observe missing values for UPC, FileLineNumber. Which cannot be replaced. And hence we decided to drop incomplete rows. So we are left with 642925 rows of data. Since its structured data we are storing it in data frame data structure.

We are planning to undertake following potential feature engineering tasks

* Add field for weekend / workday
* Create Shopping cart category (small / medium / large )
* Categories Departments
* Identify entropy of shopping cart
* Create Shopper Profiling - we can do clustering to determine profiling
* Define Boolean value - returns / no returns
* Do Basket Analysis modelling to get additional features.

## **HIGH LEVEL ANALYTIC PROBLEM (Question C)**

As we are tasked with categorizing shopping trip types based on the items that customers purchased and we also have labeled historical data , this problem could be solved using multi-class classification techniques which is a supervised learning technique..

# **PROJECT: Part 2**

**Part 2: Model construction and evaluation**

**a)      Construct analytic model(s) to address the project objective**

**b)      Evaluate the model outcomes**

**c)       Iterate and improve the model when necessary**

**d)      Justify the final model and its output**

During the second phase of the project team members were attempting to build different models to produce the highest accuracy. Feature engineering yielded several additional dimensions added to the data set.

Aleksey was trying to build a tree adding and testing dimensions to the data set, but get caught up in mangling with data more than concentrating on the actual model. The model Aleksey produced roughly yields 48% accuracy, which is significantly lower than the model that was built by Jagger.

Jagger worked to converted Department Description column to multiple binary columns. As well as worked to create new columns such as basket\_item\_count, basket\_return\_count , basket\_size. And then merged rows corresponding to each basket / Visit into single row , so that this data could be used for classification.

Jagger utilized Random Forest algorithm to produce a model that is almost 66% accurate. Jagger used ensembling approach to test accuracy of the different types of models. Random Forest approach yielded best results. As such, the Random Forest model produced by Jagger is a clear winner. In addition, Jagger also ranked the model following Kaggle’s guidelines – Jagger script contains the details. Our team received a Kaggle Score of 1.65507.

Aleksey’s and Jagger’s scripts can be found in the appendix portion of this document.

Winston is working on an ensemble model that offers a "majority rules" prediction based on the outputs of Jagger's random forest model, a support vector machine model, and an XGBoost model. The support vector machine model, on its own, achieved 70% precision in testing.

We will be also exploring ways to extract / create new useful features from Upc and FinelineNumber columns to further improve our accuracy.

# **PROJECT: Part 3**

# **REFERENCES / ATTACHMENTS**

## **PROJECT Part 1 Attachments**

1. EXPLORATORY DATA ANALYSIS R SCRIPTS (TEAM) 
2. EXPLORATORY DATA ANALYSIS - OUTPUT IN PDF FORMAT (TEAM) 

## **Project Part 2 Attachments**

1. Classification attempt using rpart package
2. Random forest implementation
3. Data Files: 