ISyE 7406 Project Proposal
Walmart Trip Type Classification

Xianghui Gu (xgu72)
Jiaye Liu (jliu658)
Chenwei Yue (cyue8)
Yang Wu (ywu613)

I. Project description:

Walmart spares no effort to enhance customer's shopping experience, no matter online or offline. During their observation, they find store visits of customers can be segmented into different trip types. For example, 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. This classification benefits Walmart to predict the peak of store visits and recommend suitable products to customers.

In this project, we would like to help Walmart classify the customer trip type for each visit number. The predictor features are displayed and explained in *Data Source*. We will train the model using the partial training data, select the best model using the rest of training data, and compare the testing result using testing dataset with the online true classification.

II. Data source:

In this project, we will categorize shopping trip types based on the items that customers purchased. Walmart has categorized the trips contained in the 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.

The training and testing data is provided by Walmart on Kaggle (https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data) . The training set (train.csv) contains a large number of customer visits with the TripType included and we predict the TripType for each customer visit in the test set (test.csv). Each visit may only have one TripType.

- TripType - a categorical id representing the type of shopping trip the customer made. This is the ground truth that we are predicting. TripType_999 is an "other" category.
- VisitNumber - an id corresponding to a single trip by a single customer

- Weekday - the weekday of the trip
- Upc - the UPC number of the product purchased
- ScanCount - the number of the given item that was purchased. A negative value indicates a product return.
- DepartmentDescription - a high-level description of the item's department
- FinelineNumber - a more refined category for each of the products, created by Walmart

For the test set, we will predict all candidate trip types for each visit number, and a probability for each class. Results are evaluated using the multi-class logarithmic loss. For each visit, a set of predicted probabilities is computed (one for every trip type). The formula is:

$$-\frac{1}{N}\Sigma_i\Sigma_j y_{ij} log(p_{ij})$$

where *N* is the number of visits in the test set, *M* is the number of trip types, log is the natural logarithm, y_{ij} is 1 if observation *i* is of class *j* and 0 otherwise, and p_{ij} is the predicted probability that observation *i* belongs to class *j*.

III. Scientific Research Questions

1. What is the impact of this project towards Walmart?
2. Is there any similar study done before? What are their approaches?
3. What's new in your approach and why do you think it will be successful?
4. What are the risks and the payoffs?
5. How long will it take?
6. Are there any missing values in the training data set? How to deal with them?

IV. Proposed Methods

1. Logistic Regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Its advantages include:

- It does not assume a linear relationship between the independent variables and dependent variable, so it may handle nonlinear effects.
- The independent variables don't have to be normally distributed, or have equal variance in each group.
- We can add explicit interaction and power terms to improve the accuracy of our models.

2. Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation.
- Able to handle both numerical and categorical data.

3. Random Forest

We also consider an ensemble method, random forest (RF). RF is an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. The advantages of RFs are:

- It can handle thousands of input variables without variable deletion.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- Prototypes are computed that give information about the relation between the variables and the classification.