OVERVIEW OF DATA INFORMATICS IN LARGE DATA ENVIRONMENTS

PROJECT

TRIP TYPE CLASSIFICATION

BY

SURESH KASIPANDY AND NIKHIL SHARMA

PROJECT DEFINITION [1]

Walmart attempts to provide every customer the best shopping experience they can through segmentation of stores visits into different trip types. These trip types are created using a combination of “art” (preexisting customer insights) and “science” (historical purchase data). In this project, we aim to create a classification model and predict the trip types by focusing only on the data (“science”) using a training dataset available on Kaggle containing a much smaller set of features than what is used by Walmart. Walmart posted this project idea/challenge on Kaggle as they believe that their segmentation process would see refinement by improving the science behind trip type classification.

BACKGROUND

Machine learning is a field of computer science that deals with the creation and study of algorithms that can teach a computer to learn from pre-existing data sets and use the knowledge obtained to make future predictions. It is a study that saw evolution through collaborations between computational learning theory related to artificial intelligence and pattern recognition. There are several approaches that can be applied in Machine Learning depending on the nature of the problem or dataset being used. Some of the more notable approaches are Naïve Bayes Classifier, Decision Trees, Nearest Neighbor, Support Vector Machines and Neural Networks. These approaches can be used to solve different kinds of problems such as classification, regression and clustering. Machine Learning can be applied in a large variety of fields such as advertising, bioinformatics, fraud detection, recommender systems and stock market analysis. The list of markets that employ the use of Machine Learning and Data Science is constantly growing.

For the sake of this project, we aim to create a classifier algorithm to identify trip types in the Walmart dataset. In particular, the approach decided upon was Decision Tree as it was deemed most suitable for the nature of the problem. In decision tree, observations of an item are mapped to conclusions of a target value related to the item. This model map is then used to predict the class of future item inputs.

DESCRIPTION OF DATASET [1]

Walmart has classified their transactions into one of 38 different trip types using their own proprietary classification method. The provided dataset has a more limited set of features that are to be used to develop our own classification/clustering method. The training dataset contains 684672 observations and is a CSV file of size 30.9 MB. The Test dataset contains 653646 observations and is a CSV file of size 29.6MB. . The provided features are:

- **TripType** - a categorical id representing the type of shopping trip the customer made. This is the ground truth that you 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

PROJECT PLAN

This project requires supervised learning/classification technique to construct a model based on the training dataset. Decision trees will be used to classify the test data according to the specific feature values that become increasingly specific. 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 system will be designed in Python using the Jetbrains Pycharm code editor.

WORK DIVISION

TASK	**TEAM MEMBER**
Data Preprocessing	Suresh Kasipandy
Data Classification	Nikhil Sharma
Testing	Nikhil Sharma
Documentation	Suresh Kasipandy

METHODS

a) **TOOLS**

i. *Scikit-learn*[3]:

Scikit-Learn is a machine learning library for use with the programming language known as Python. It is open source and is very versatile allowing for advanced analysis on data. It features various Machine Learning algorithms including

Decision Trees, Random Forest, Support Vector Machines, k-means, Gradient Boosting and DBSCAN. The package was initially released in 2007 and has since seen use by many large companies in big industries.

ii. *Pandas*[4]:

Similar to scikit-learn, Pandas is an open source library for performing data analysis using Python. It provides easy-to-use, simple, high performance data analysis tools and data structures for Python.

In this project, we are using pandas to load data from CSV files.

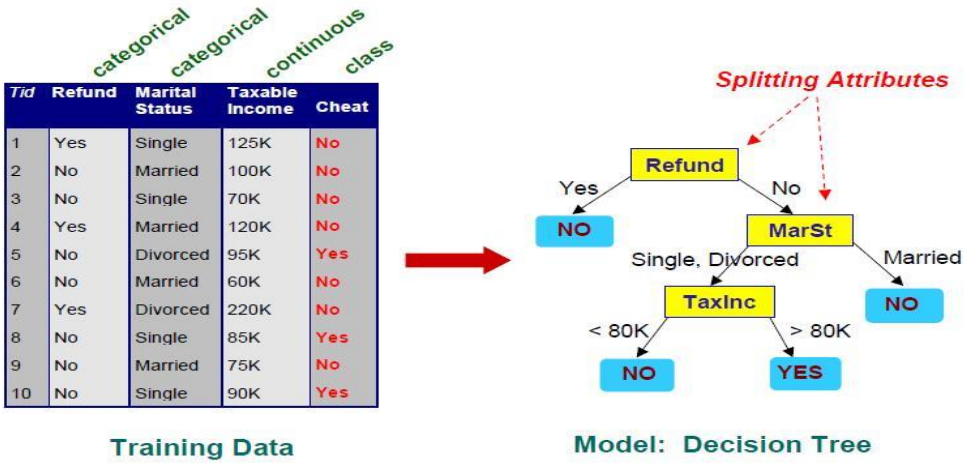
iii. *Jetbrains Pycharm code editor*:

Pycharm is an Integrated Development Environment (IDE) for creating programs using the Python programming language. It consists of a variety of tools including a graphical debugger, an integrated unit tester and code analysis among others.

b) Algorithm:

DECISION TREE [6]

A decision tree is a classification tree which is a synopsis of the classification scheme. It represents a hierarchical structure containing descriptors for classes or groups based on which predictive analysis is conducted for future inputs. Within the structure, items are constantly separated into classes and subclasses based on attributes until further splitting is not possible.



EXPERIMENT

In This project, we use DecisionTreeClassifier class from scikit-learn package capable of performing multi-class classification on a dataset.

We have read the training data provided by Walmart using pandas dataframe. The string valued columns ‘Weekday’ and ‘Department Description’ have been encoded to float numbers before we fit the data.

DecisionTreeClassifier takes as input two arrays: an array X, sparse or dense, of size[n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples [3].

The decision tree is fitted using the DecisionTreeClassifier inbuilt class dt.fit(X, y) where X is the list of attributes we use for classification and y is the target class attribute (TripType).

The test data is imported using pandas dataframe and the class for each item in the dataset is predicted using the DecisionTreeClassifier inbuilt class dt.predict(Data), where Data is the data from the columns used for prediction.

The final output with the predicted TripType classification for the test data has been included in the final_output.csv file attached in the Final_Project.zip file. The first column in the final_output.csv file gives us the predicted TripType classification.

Performance Analysis and Results:

Performance analysis with varied size of Training data:

The performance analysis of the project with different values of the training dataset is as shown below. However, the size of the test data which is 653646 is kept the same.

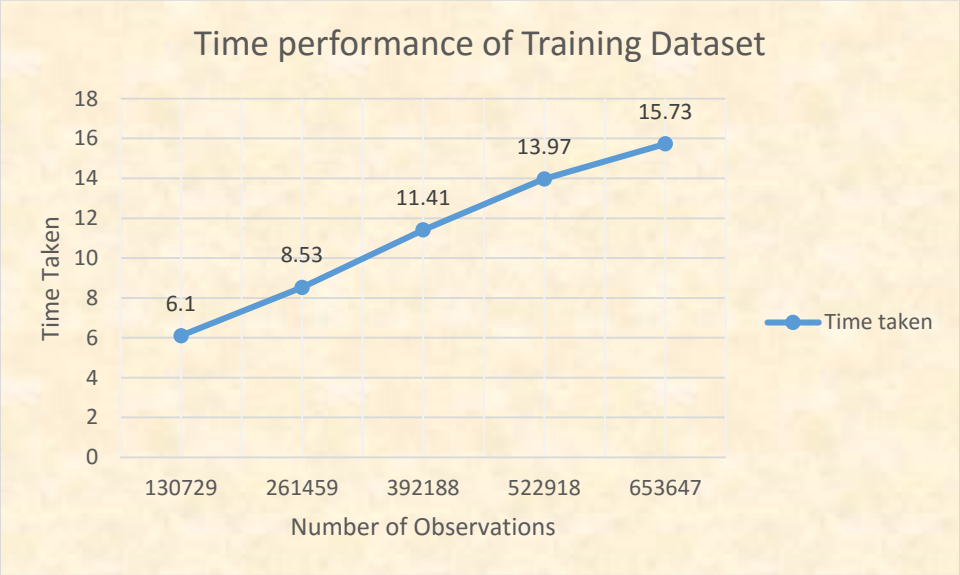


Chart1. The running time of Decision Tree Classifier with varied size of training data

CONCLUSION

For years retailers have employed atmospherics; considering the physical environment and how it affects consumer shopping behavior and how they can be leveraged to optimize the retail experience for the consumer. To locate individual or groups of products based on transaction groupings is an interesting exercise for both increasing customer satisfaction and revenue

This project will help enhance product placement and product assortment, promo displays, understand what type of customers/motivations a store layout or a store type or channel type is addressing and to determine if it is as per the its goals and assess if there is a change in shopping motivations compared to in the past.

Our project has provided Walmart with a classification of the test data into various Trip Type as required. Some of the categories of shopping trip types could be a) Daily shop b) Urgent item c) Fill in d) Stock up or e) Special occasion f) Recreation/fun shopping g) or to buy ready to eat items.

Walmart can assess things such as when consumers on a particular shopping trip type (for example, Stock up trip) are more likely to try a free sample item and purchase it.

Using our classification model, the Trip types have been predicted using the given data by Walmart.

REFERENCES

1) The project has been posted as a competition on Kaggle.

   https://www.kaggle.com/c/walmart-recruiting-trip-type-classification

2) Ethem Alpaydın, Introduction to Machine Learning, Second Edition, MIT Press, 2010.

   http://www.realtechsupport.org/UB/MRIII/papers/MachineLearning/Alppaydin_Machine Learning_2010.pdf

3) Scikit-learn Decision tree documentation.

http://scikit-learn.org/stable/modules/tree.html

4) Pandas python Data analysis library.

http://pandas.pydata.org/

5) Jetbrains Pycharm code editor:

https://www.jetbrains.com/

6) Decision Tree classifier.

http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html