REPORT

Predicting products purchase in the next order

Muhammad Umar Salman 20100120

Momina Haider 20100268

Razi ul Haq 20100015

To predict the purchases in the next order we used 3 different methods to predict which values will be next purchased.

The following methods are:

1. Logistic Regression
2. Clustering
3. Decision Trees

Using each of these 3 methods by implementing them on python through different python libraries such as

* from sklearn.cluster import KMeans
* from sklearn import linear\_model
* from sklearn.tree import DecisionTreeClassifier

for each of the three methods.

DECISION TREES

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

A decision tree classifier is just like a flowchart diagram with the terminal nodes representing classification outputs/decisions. Starting with a dataset, you can measure the entropy to find a way to split the set until all the data belongs to the same class. There are several approaches to decision trees like ID3, C4.5, CART and many more. For splitting nominal valued datasets you can use the ID3 algorithm. You can use matplotlib library to visualize the tree data. Decision Trees are prone to overfitting, thus to avoid overfitting you can prune the decision tree by combining the adjacent nodes that have low information gain.

CLUSTERING

Clustering is the grouping of a set of objects based on their characteristics, aggregating them according to their similarities. Regarding data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. On the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, differentiating its properties and results. These models are distinguished by their organization and type of relationship between them. The most important ones are:

**– Centralized**– each cluster is represented by a single vector mean, and an object value is compared to these mean values  
**– Distributed** – the cluster is built using statistical distributions  
**– Connectivity** – he connectivity on these models is based on a distance function between elements  
**– Group** – algorithms have only group information  
**– Graph** – cluster organization and relationship between members is defined by a graph linked structure  
**– Density** – members of the cluster are grouped by regions where observations are dense and similar

LOGISTIC REGRESSION

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Sometimes logistic regressions are difficult to interpret; the Intellectus Statistics tool easily allows you to conduct the analysis, then in plain English interprets the output.

*Through these methods we have implemented the predicted values whether a user will next purchase what item. This is done by testing the data and predicting by making the label “Reordered”. And to train the data we use all the numerically significant values which may give us some sort of pattern to get an accurate result of what a particular user will use next. The users and the products are printed after running the code run.py for each of the three methods.*