**Prediction of customer repeat purchase patterns using ensemble classification and feature engineering**

**Team: FP - 8**

**Group Members:**

Annu Sharma (36731206)

Srinivas Gubbala (21317376)

Venkat Trived Menta (87371010)

Keke Zhai (98697900)

Nicholas Troiano (69547964)

Kishan Rao (68568951)

**Problem Statement:**

Using historical transaction records of customers collected over a year, we try to predict the ‘repeat purchase’ behavior of an existing customer employing an ensemble classification model trained on the repeat purchase behaviors of a subset of customers. The problem statement is inspired from the Acquire Valued Shoppers problem hosted at Kaggle which has most commonly been approached using Quantile Regression and Random Forest Models of which the state-of-the-art model achieves an accuracy of ~57%.

**Novelty:**

We attempt to innovatively engineer the feature set for the problem and experiment with ensemble models including Collaborative Filtering Algorithms and Logistic Regression with varying regularisation and bagging to boost the performance of the classifier.

**Why it will be an interesting DS project:**

Companies spend a large capital to bring customers on board. It is equally important to retain the acquired customers to realize the value spent in acquisition and to make profits. Targeting the right customers based on the knowledge obtained from their transactional data gives incentives which not only increases the profitability but also helps in achieving a loyal customer base.

This is one of the biggest transactional dataset publicly available and provides around 350 million transactions over a period of one year and customer reactions (repeat purchase or not) to an offered coupon henceforth. Along with the given features, there is vast scope (practically the main challenge of the problem) to generate relevant features to train the classifier and hence predict better. Moreover, the problem shows a lot of variation in prediction results when the prediction models are varied, based on the literature survey.

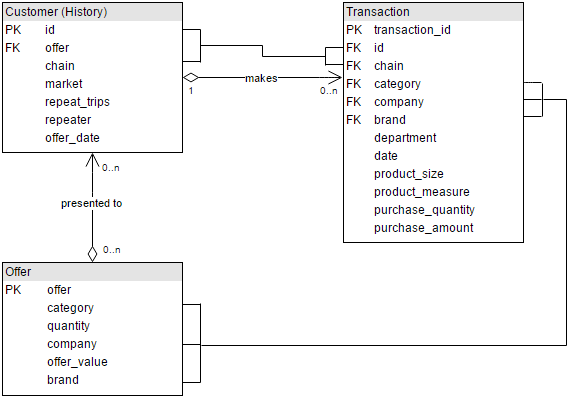
**Data Sources and Characteristics:**

We have three input sources/tables for the model(s):

**TRANSACTIONS -** It contains the transaction history of all customers (300,000) at least a year before they were offered a discount coupon. It contains around 350 million records. (19.8 Gb)

**HISTORY -** It contains information on the coupon offered to the client, and his reaction (repeat buy or not, if yes - number of times) It has information on 160,000 offers. (5 Mb)

**OFFERS -** It contains information on the pool of valid discount coupons. It has 37 items.(2 Kb)

The attributes of the three tables and their relationship is captured in the ER diagram below:

**Data Science pipeline:**

The pipeline encompasses the following steps, in order of mention.

1. **Exploratory Analysis of Data:** Identifying general trends and patterns in customer purchase history and visualising them.
2. **Cleaning Task:** Filtering the rows that are irrelevant such as transactions with purchase amount or quantity as zero, transactions with missing fields like brand or company etc.
3. **Data Reduction**: Finding the appropriate measures to create a subset of the data on the basis of the coverage of the valid coupons, say category and brand.
4. **Feature Engineering:** Designing and generating new, innovative customer behavioral features and coupon targeted products features.
5. **Feature Selection**: Identifying the most predictive features amongst the engineered features which are used as input for classifiers.
6. **Prediction:** Predicting the repeat purchase behavior of the test set of customers given the offer presented.
7. **Visualization:** Visualising the prediction results.

**Challenges:**

* Model prediction accuracy primarily depends on the quality of the extracted features in the data and the choice(s) of prediction models. Features pool are of the form - number of times customer purchased the items of a particular brand/company/category in X days before the coupon was offered, if the product returned previously, seasonal buying pattern of products, total spending by a customer on the specific brand and so on.
* Dimensionality Reduction: Finding relevant features from the engineered feature set for the prediction task to optimise the area under the ROC curve.
* Finding the right ensemble of predictors to improve upon set baseline.

**Tentative Division of Labor:**

Data Set Exploration - Venkata Trived Menta, Nicholas Troiano

Data Cleaning/Integration - Kishan Rao, Srinivas Gubbala

Feature Engineering – Srinivas Gubbala, Annu Sharma, Keke Zhai

Prediction - Annu Sharma, Kishan Rao, Venkata Trived Menta

Data Visualization - Keke Zhai, Nicholas Troiano

**End Data Product:**

The end product will be a classifier capable of learning repeat purchase behaviours of customers from past transactional patterns which can be deployed for targeted offers disbursement to both - optimise the revenue of the company as well increase customer loyalty for a brand/product.

**Related Work and Baseline:**

**Current Literature:**

Data Reduction, Feature Engineering and developing a Regression Model were the major subtasks undertaken amongst which Feature Engineering was the most pronounced and debated one. The end results varied based on the feature set that is selected to train the regression model. Different Regression models were used, prominently - Logistic Regression and Quantile Regression.

**Baseline:**

The baseline for our predictions would be the predictions obtained without the newly engineered features i.e using only the product features provided by the datasets. We will build on this accuracy by engineering and testing new features (customer, product features) on the ensemble model.

**Measure of Success:**

Our prediction results will be tested against the ground truth provided by Kaggle which will be evaluated on the area under the ROC (Relative Operating characteristic Curve) which compares the two operating characteristics - True Positive Rate and False Positive Rate. We also plan to use K-Folds Cross Validation and F-Measure metrics to evaluate the model performance.