Problem Statement

- Inspired by the ‘Aquire Valued Shoppers’ problem hosted on Kaggle.
- Use historical transaction records of customers for over a year to predict ‘repeat transaction’ behavior of customers.
- Attempt to innovatively enhance the feature set of the problem.
- Use ensemble model which includes Collaborative Filtering and Logistic Regression with varying degrees of regularisation.

Why It is an interesting DS project

- Will help companies which spend a large amount of capital on acquiring customers to retain them and increase profitability.
- One of the largest transactional datasets publicly available.
- Record of 350 million transactions over a period of a year.
- Vast scope to generate additional features to train the classifier and improve the prediction results.

Description of data

Transactions Table		
Column ID	**Description**	**Foreign Key**
ID	A unique id representing a customer	Maps with id in History table
Chain	An integer representing a store chain	Maps with chain in History table
Department	An aggregate grouping of the Category	
Category	The product category	Maps with category in Offers table

Column Id	Description	Foreign key
Company	An id of the company that sells the item	Maps with company in Offers table
Brand	An id of the brand to which the item belongs	Maps with brand in Offers table
Date	The date of purchase	
product_size	The amount of the product purchase	
product_measure	The units of the product purchase	
product_quantity	The number of units purchased	
product_amount	The dollar amount of the purchase	

Offers table		
Column ID	**Description**	**Foreign Key**
Offer	An id representing a certain offer	
Category	The product category	
Quantity	The number of units one must purchase to get the discount	
Company	An id of the company that sells the item	
offer_value	The dollar value of the offer	
brand	An id of the brand to which the item belongs	

History Table		
Column ID	**Description**	**Foreign Key**
ID	A unique id representing a customer	
Chain	An integer representing a store chain	
Offer	An id representing a certain offer	Maps with offer in Offers table
Market	An id representing a geographical region	
repeat_trips	The number of times the customer made a repeat purchase	
Repeater	A Boolean, equal to repeat_trips > 0	

Pipeline/Workflow and Components

- **Cleaning Task:** Filtering the junk and irrelevant rows.
- **Integration/Aggregation/Association/Compression:** preparing rectangular matrix with each customer as a row
- **Prediction (Analysis):** To model patterns in the customer to company/customer to product relationship
- **Visualization:** Presenting a visual representation of trends in customer behavior.

Challenges

- Most of the model is based on extracted features of data such as the number of times a customer purchased an item in a certain number of days before a coupon was offered. This requires significant pre processing of data before it can be made usable for cleaning/prediction.
- **Dimensionality Reduction :** Finding relevant features from the engineered feature set for prediction task.
- Varying the offer categories for training and test data.

Division of Labor

1. Dataset analyzation/visualisation
2. Brainstorming of problem statement
3. Current Literature
4. Proposal writing
5. Data Cleaning
6. Data Integration/Compression
7. Data Analysis (Prediction)
8. Data Visualization
9. Report Writing
10. Poster Preparation

End Data Product

- A model which will predict customer behavior.
- A deeper understanding of factors involved in product purchase.

Related Work and Baseline

- Generation of more features to predict customer loyalty from the current features.
- eg:purchased_category_brand_company

Measure of Success/Ground truth

- Prediction Results will be tested against ground truth provided by Kaggle.
- Evaluated on the area under the ROC(Relative Operating Characteristic Curve) which compares True Positive Rate and False Positive Rate.
- We plan to use K-fold cross validation and the F-measure to evaluate model performance.