# Capstone Proposal: Predicting a Basket’s Customer Segment

## Domain Background

Customer segmentation is considered to be a cornerstone of personalized marketing. Much work has been done in taking historical purchases and building recommender systems; indeed, the success of companies such as Amazon has been credited to such systems. Recommenders are only as strong as the data they’re built on.

Professionally, I’m tasked with teasing out the impact of promotions to better allocated limited promotional budget to win the hearts and wallets of particular customer segments, without alienating others. Therefore, I have a vested interest in getting as complete a picture as possible, and where data is missing, finding the best possible value for the missing field.

## Problem Statement

For brick and mortar retailers, the available data is neither as rich nor as complete as it may be for purely online retailers: full customer profiles are only available if the retailer has a solid loyalty program, and rarely include information that could be derived from things like referring websites and site searches. Even with a strong loyalty program, a retailer will find that there are many cash purchases, or purchases outside of the loyalty program where little is known about the customer. These non-loyalty transactions can represent a significant portion of any dataset and, therefore, any trends found in the rich subset may be entirely wiped out by noise from the sparse.

## Datasets and Inputs

The dataset I will use for this project is the “50,000 customers” subset of DunnHumby’s “Let’s Get Sort-of-Real: Dummy Data to Test Techniques and Algorithms” dataset, available at <https://www.dunnhumby.com/sourcefiles>

There are two tables provided: a time table with calendar information, and the main table whose columns can be divided into five sections:

* Time (week, date, weekday, and hour of the transaction)
* Basket (basket ID, quantity and dollars spent of each item, two basket size classifications, basket price-sensitivity classification, and a rough basket driver)
* Product (product code and hour levels of product hierarchy)
* Customer (unique customer code, customer’s price sensitivity, life stage)
* Store (store code, size and location)

## Solution Statement

For this capstone project, I will take a set of transactional data from fictitious retailer, graciously provided by DunnHumby and use it to simulate the above situation, where customer information is not known about certain transactions, namely their two variables “customer life-stage” and “customer price sensitivity.” Specifically, I will remove these two fields from a testing subset of the data and build a predictive model using other factors in the dataset such as product hierarchy, shop time/day, etc. While, in my day job, I have access to a much richer dataset, I believe this will act as a solid proof-of-concept to show the value of a fully-realized system.

Note that, I may choose to exclude basket price sensitivity from the input data as it may unduly simplify the solution.

First, I will investigate ways of aggregating transactional data into a per-basket profile; this could be as simple as using spend-per-department as a set of 90 variables, though I intend to experiment with dimensionality reduction. I will test a variety of classification algorithms on the data, because I have no pre-existing preference about which ones to use. I will place emphasis on ensemble methods and artificial neural networks, because I believe this will be a complex problem (e.g. I don’t expect to see linear separability or to be able to come up with a kernel function that would make Support Vector Machines a viable choice). I will also test out decision trees because of the simplicity of their explanation.

## Benchmark Model

Since we have a known testing dataset with correct answers, any solution will be graded against this “benchmark.” However, a further hurdle will be set based on two simple decision trees that sort baskets into the 4 customer price sensitivities and 7 life stages (resulting in 28 target values). Red indicates a leaf of the decision tree.

* Price sensitivities: LA=Less Affluent, MM=Mid Market, UM=Up Market, XX=unclassified
* Life stages: YA=Young Adults, OA=Older Adults, YF=Young Families, OF=Older Families, PE=Pensioners, OT=Other, XX=unclassified

## Note that “unclassified” and “other” sensitivities are intentionally excluded, as I have no insight into those groups at this time.

## Evaluation Metrics

I will be considering two key metrics in evaluating this project:

* Classification accuracy:
* Precision/Recall for each of the 28 target values:
  + Precision:
  + Recall:

Classification accuracy is useful because it’s a single, easy-to-understand metric for overall model performance. Precision/Recall for each class important because, in the real world, I care more about some classes than others (e.g. a maternity store would not care if it misclassifies a pensioner as an older family nearly as much as if it misclassifies a young adult as a pensioner).

I believe that this project is feasible, because I was able to obtain classification accuracy of 67% with an un-tuned kNeighbours classifier while exploring the data to write this proposal.

# Capstone Proposal: Section Goals/Rubric

## Domain Background

*Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.*

## Problem Statement

*Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.*

## Datasets and Inputs

*The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.*

## Solution Statement

*Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.*

## Benchmark Model

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

## Evaluation Metrics

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

## Project Design

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

## Presentation

Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.