**Data Files**

| **File Name** | **Available Formats** |
| --- | --- |
| sampleSubmission | [.csv (924.99 kb)](https://www.kaggle.com/c/allstate-purchase-prediction-challenge/download/sampleSubmission.csv) |
| train.csv | [.zip (5.63 mb)](https://www.kaggle.com/c/allstate-purchase-prediction-challenge/download/train.csv.zip) |
| test\_v2.csv | [.zip (2.02 mb)](https://www.kaggle.com/c/allstate-purchase-prediction-challenge/download/test_v2.csv.zip) |

**Files**

The training and test sets contain transaction history for customers that ended up purchasing a policy. For each customer\_ID, you are given their quote history. In the training set you have the entire quote history, the last row of which contains the coverage options they purchased. In the test set, you have only a partial history of the quotes and do not have the purchased coverage options. These are truncated to certain lengths to simulate making predictions with less history (higher uncertainty) or more history (lower uncertainty).

For each customer\_ID in the test set, you must predict the seven coverage options they end up purchasing.

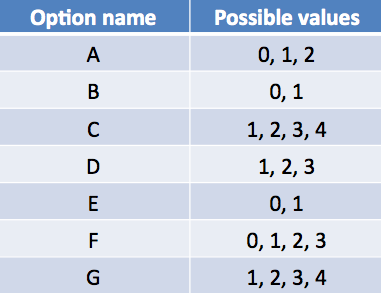
**What is a customer?**

Each customer has many shopping points, where a shopping point is defined by a customer with certain characteristics viewing a product and its associated cost at a particular time.

* Some customer characteristics may change over time (e.g. as the customer changes or provides new information), and the cost depends on both the product and the customer characteristics.
* A customer may represent a collection of people, as policies can cover more than one person.
* A customer may purchase a product that was not viewed!

**Product Options**

Each product has 7 customizable options selected by customers, each with 2, 3, or 4 ordinal values possible:



A product is simply a vector with length 7 whose values are chosen from each of the options listed above. The cost of a product is a function of both the product options and customer characteristics.

**Variable Descriptions**

**customer\_ID** - A unique identifier for the customer  
**shopping\_pt** - Unique identifier for the shopping point of a given customer  
**record\_type** - 0=shopping point, 1=purchase point  
**day** - Day of the week (0-6, 0=Monday)  
**time** - Time of day (HH:MM)  
**state** - State where shopping point occurred  
**location** - Location ID where shopping point occurred (10001,10002,…16580,NA)  
**group\_size** - How many people will be covered under the policy (1, 2, 3 or 4)  
**homeowner** - Whether the customer owns a home or not (0=no, 1=yes)  
**car\_age** - Age of the customer’s car  
**car\_value** - How valuable was the customer’s car when new ("a","b",…"i",NA)  
**risk\_factor** - An ordinal assessment of how risky the customer is (1, 2, 3, 4, NA)  
**age\_oldest** - Age of the oldest person in customer's group  
**age\_youngest** - Age of the youngest person in customer’s group  
**married\_couple** - Does the customer group contain a married couple (0=no, 1=yes)  
**C\_previous** - What the customer formerly had or currently has for product option C (~~0=nothing~~, 1, 2, 3,4, NA)  
**duration\_previous** - how long (in years) the customer was covered by their previous issuer  
(0,1, 2, …15, NA)  
**A,B,C,D,E,F,G** - the coverage options

**cost** - cost of the quoted coverage options

## Evaluation

Submissions are evaluated on an all-or-none accuracy basis. You must predict every coverage option correctly to receive credit for a given customer. Your score is the percent of customers for whom you predict the exact purchased policy.

## Submission File

The submission format is created by concatenating each plan option (A,B,C,D,E,F,G) as a single string, in order. The file should contain a header and have the following format:

customer\_ID,plan  
10000001,1111111  
10000002,1111111  
10000003,1111111  
...

### COMPETITION-SPECIFIC TERMS

COMPETITION NAME (the 'Competition'): Allstate Purchase Prediction Challenge

COMPETITION SPONSOR: Allstate

COMPETITION WEBSITE: https://www.kaggle.com/c/allstate-purchase-prediction-challenge

PRIZES: $50,000

WINNER LICENSE TYPE: Non-exclusive

MAXIMUM TEAM SIZE: as listed above

MAXIMUM SUBMISSIONS PER DAY: as listed above

## Timeline

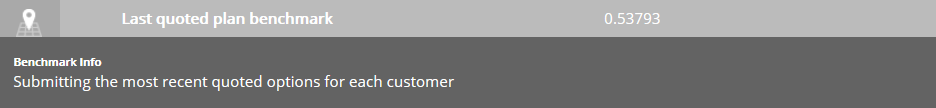
## Competition

* **May 12, 2014** - Deadline for new entrants and team mergers
* **May 19, 2014** - Final submission deadline

All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The organizers reserve the right to update the timeline as the competition progresses.

## Leader Board





**Data**

|  |  |  |
| --- | --- | --- |
| **File** | **Expanded** | **Rows** |
| train.csv | 46MB | 665,249 |
| test\_v2.csv | 14MB |  |

In Excel it takes 243MB RAM to open train.csv

|  |  |  |
| --- | --- | --- |
| **Column** | **Domain** | **Missing** |
| customer\_ID | 97,009 Customers 665,249 Shopping Pt |  |
| shopping\_pt | 1,2,…13 | No |
| record\_type | 0,1 | No |
| day | 0,1,…6 | No |
| time | 00:01…23:59 | No |
| state | "AL", …"WY" | No |
| location | 10001,10002,…16580 | No |
| group\_size | 1,2,3,4 | No |
| homeowner | 0,1 | No |
| car\_age | 0,1,2,…85 | No |
| car\_value | "a","b",…"i" | No |
| risk\_factor | 1,2,3,4,NA | Yes |
| age\_oldest | 18,19,…75 | No |
| age\_youngest | 16,17,…75 | No |
| married\_couple | 0,1 | No |
| C\_previous | 1,2,3,4,NA | Yes |
| duration\_previous | 0,1,2,…15,NA | Yes |
| A | 0,1,2 | No |
| B | 0,1 | No |
| C | 1,2,3,4 | No |
| D | 1,2,3 | No |
| E | 0,1 | No |
| F | 0,1,2,3 | No |
| G | 1,2,3,4 | No |