**This section of the site contains information confidential to Blue Martini Software and Gazelle.com Inc.**
**Use of the data is restricted by a non-disclosure agreement that you must have signed at http://www.ecn.purdue.edu/KDDCUP/**

# KDD Cup 2000 FAQ

1. 5/27/00 The problem is really hard! I can barely get the accuracy down. Am I doing something wrong?

   The problems are really hard. In fact, many algorithms will do worse than predicting the majority class.
   However, getting even a small improvement may be significant, especially if some predictions are high confidence.
   For example:
   - It is very hard to guess whether a session will end or what a user will browse given a single page view.

     However, perhaps it is possible to make such predictions for sessions of length 5 or more.
   - There may be attributes that need to be created. With these attributes, the problem may be easier.
2. 5/23/00 In the names file, some fields are marked as "ignore." What does this field type mean and are we allowed to use it?

   The names file format we supplied lists all the values for a column. When there are many values (over 100 distinct) or when there is a single value, we mark the column as ignore. Such attributes have very little value for data mining algorithms like C4.5 or C5.0. You may want to construct higher level features with fewer values, or you may have an algorithm that can deal better than C4.5/C5.0 with such attributes.
   Other attribute types are:
   - continuous: the values are either integer or float
   - date: date values in the format of "yyyy-MM-dd"
   - time: time values in the format of "HH\:mm\:ss"
   - a list of discrete values: unordered discrete values
   - [ordered] + a list of discrete values: ordered discrete values (the order of the values makes semantic sense).

   More information about the C5.0 format can be found in http://www.rulequest.com/see5-win.html at the RuleQuest web site. Note that the available version of C5.0 does not support "time" (we added this ourselves).
3. 5/26/00 Some sessions in the Clicks dataset have a very large numbers of pages (in thousands). How can this be explained?

   Welcome to the WWW where bots and crawlers visit. Several thousand hits per session are common for crawlers. We suggest you look at the "User agent" column, which the browser sends as its identification. Sometimes it will disclose itself as a crawler; other times it will have a slightly different variant of a true browser, and at times it will just hide itself as any other browser. Some examples are given in http://mosa.unity.ncsu.edu/~brabec/antispam.html
4. 5/26/00 Are the non-Acxiom demographic fields (things likeYourFavoriteLegcareBrand) self-reported?

   Yes. The questions were asked during registration. Note, however, that (as is very typical of new sites), the registration questions changed several times during the period in which the data was logged, and hence some customers who registered have values for one set of questions and other customers who registered have values for another set of questions.
5. 5/26/00 Will Acxiom demographic data be part of the test data set as well?

Yes. However, you must realize that Acxiom demographics only exist for users who have registered (and given their name/address). For the questions with a test set (1,2,4,5), this represents a very small portion of the visitors. Also, a registered user may come back to the site and not login (cookie tricks may help you match, but there are few repeat visitors in the short duration we have).

6. 5/26/00 Does the value of '?' indicate a missing value (NULL) as it does in C4.5 and C5.0?

   Yes. Note, however, that based on some feedback about unknown versus inapplicable, the updated dataset to be released 5/27/00 will specify NULL instead of "?" for booleans. The unknowns (?) will still exist for categorical where the semantics is more appropriate.
7. 5/26/00 Is the combination of "Session ID" and "Request Sequence" a unique identifier for every record?

   For the clicks dataset, it is.
8. 5/26/00 Is the attribute "Order Line Session ID" in the clicks dataset the same as "Session ID" in the order dataset?

   Yes! Most columns that have the same name (e.g., user id) will match across the two tables.
9. 5/26/00 Should the answer to Content Question #5 be provided in terms of visitors (demographics), in terms of sessions, or either?

   Either, i.e., any insight that would interest a business user is reasonable.
10. 5/27/00 Viewed Brand in the clicks dataset has several values. What is NULL?

    Null means that the page view is not a product details page.
11. 5/27/00 Order Amount is sometimes zero or a small negative number. What does it mean?

    Zero purchases are possible with coupons and Gaezlle had a $10 off coupon promotion with no minimum price.
    We're asking Gazelle to explain the small negative amounts.
12. 5/28/00 Could you be as specific as possible in describing the differences between the web site during the period Jan 30, 2000-April 1, 2000 and the site today?

    One of the hallmarks of the Blue Martini system is the ability to change something and stage a new site while it is up and running. Gazelle has staged 190 versions of the site since they started building the site. Fifty five (55) of those were done from Jan 30 to April 1st, so we could not easily describe all the changes. The most obvious changes were:
    - Change in the registration form. As you can see, some customer attributes are null after a certain period, indicating they were not asked any more, while others started getting values late.
    - New promotions. The common ones are described in the introduction document.
    - New product introductions. Over time, the number of products and brands grew significantly.

      Our system has a notion of an object (products, product attribute value, template, promotion, basically any object). On Jan 30, there were 6,262 object; on Feb 15 there were 11,742; on Feb 28 there were 14,767; on March 15 there were 16,146; and on March 31 there were 17,281. As you can see, the site is live and growing, which makes the problem harder. The good news is that product and assortment ids that we log are stable, so when you see a product logged, it remains the same product even if templates changes are made or it gets additional attributes, etc.
13. 5/29/00 In aggregated file for question 2 there are records with several target fields marked as "True" at the same time. For example record with SessionID = 109 has following data:

    Num Hanes Viewed Later = 2
    Num AmericanEssentials Viewed Later = 1
    Num DonnaKaran Viewed Later = 0
    No Top Brand Viewed Later = False

    which means that in the remainder of the session the visitor will view 2 pages of Hanes and 1 page of AmericanEssensials. The question is:What should be predicted in such case (Hanes or AmericanEssentials) ?
    Answer: You will get the maximum, i.e., two points if you predict either Hanes or

AmericanEssentials. Given the motivation in our docs (a single link on the page to the brand) that defines the evaluation criteria, you should predict the brand with the higher confidence of a visit in the remainder of the session.

14. 5/30/00 In the answer to question 5 above, we mention that Acxiom demographics only exist for users who have registered (and given their name/address). For the questions with a test set (1,2,4,5), this represents a very small portion of the visitors. However, the given statistics show that majority of Axiom's columns have more than 50% of non-null records, such Gender. Why?

    The reason is that this statistics is computed at the customer level using Customer ID. That is, each customer with a unique Customer ID is counted once. Since all unregistered visitors have no customer ID (null), they are treated as a single customer (with null as the ID value) during this statistics calculation.
    Therefore, the number of "Null Values" here is the number of registered customers for whom Acxiom does not found Gender + 1.
15. 5/30/00 What do "Order Line" and "Order" mean?

    Each "order" can have one or more "order lines." For example, in a single order you can buy 2 pairs of socks A and 1 bottle of body cream B. This order will have two order lines -- one for socks A and one for body cream B.
    An example of an order line variable is "Order Line Quantity" which indicates the number of units ordered. In the order above this would be "2" for the socks and "1" for the cream.
    An example of an order variable is "Order Amount" which is the total amount of money spent for the entire order. This includes the amount of money spent for each order line as well as tax and shipping.
16. 5/30/00 What does the "HasDressingRoom" column mean?

    Gazelle.com has a feature, called "dressing room", that allows you to see the visual effect of a product such as when wearing a Thigh High. You can change skin tone and hosiery and color there. The attribute HasDressingRoom tells whether a product has this feature or not.
17. 5/30/00 What does LeadTime mean?

    Time in days from when a product is ordered to when it is expected to be receipted in the fulfillment center and available for sale.
18. 5/30/00 What do KS,TS,ML,OC, andPA in Socktype1 & Socktype2 stand for?

    KS: knee socks
    TS: trouser socks,
    ML: mid-length
    OC: over-the-calf
    PA: peds and ankle-highs
19. 5/30/00 What do MAS, BKS, BDCS, MCS, KP, TH, LG, MDS, TT, WAS, WDCS, FO, LEO, GDCS and PH in CategoryCode stand for?

    MAS: men's athletic socks
    BKS: bike shorts
    BDCS: boy's dress/casual socks
    MCS: men's casual socks
    KP: knee-highs/peds
    TH: thigh-highs
    LG: leggings
    MDS: men's dress socks
    TT: tights
    WAS: women's athletic socks
    WDCS: women's dress/casual socks
    FO: fashion/other
    LEO: leotards
    GDCS: girl's dress/casual socks
    PH: pantyhose
20. 5/30/00 What do CT and STW in WaistControl: CT and STW stand for?

CT: control top
STW: sheer-to-waist

21. 5/30/00 What do UBC, MBC, BS and LBC in BodyFeature stand for?

    Gazelle.com doesn't use body feature. This is used as place-holders.
22. 6/3/00 What does "Last Page" or "Last" in some aggregated column names mean?

    In the aggregated data for question 1 and question 2, it means that this column contains the value of the last request in a session for the corresponding column in the un-aggregated data. In the aggregated data for question 3, it means that this column contains the value of the last request in the last session for the corresponding column in the un-aggregated data.
    For example, "Texture Last" in q3_agg means the texture of the product viewed in the last page of the last session (it has value null if this page is not a product detail page); "Session Last Template" in q2_agg means the template of the last page of a session.
23. 6/5/00 Will the test set be in the same format as the training set? Will it be from the data collected in the .jhtml format or the .jsp format that gazelle.com is currently using?

    The test dataset is in the same format as the training set for each question if applicable. Both of them were collected in the .jhtml format.
24. 7/10/00 What is the equivalent of a static URL in the dataset?

    Template + Query string (where + means concatenate).
25. 7/10/00 Is there a unique key in the data?

    Session-id and sequence form a unique key.
26. 7/10/00 What is the meaning of gazelle.com being a referral site to itself?

    This might happen sometimes. For example, when a customer visits the gazelle.com site and stops at a place for a while. This makes the session time out. If the customer restarts the visit, a new session will be created. In this case the referral of the second session will be gazelle.com.
27. 7/10/00 At which situations should we expect the value of "session last query string" to be NULL ?

    Normally, query string contains parameters passed to the page. Some page may not have parameters. In this case, query string is null.
    "Session last query string" is the query string of the last page of the session in the data set. Note that this page may not be the ending page of the session, because we clipped some sessions when creating the datasets. This column is computed by us based on the training portion only, so there is no "leak" of the target.
28. 7/10/00 What exactly is a "Dressing Room" ?

    Could you please refer us to one product/page that has it?
    Gazelle.com has a feature, called "dressing room", that allows you to see the visual effect of a product such as when wearing a Thigh High
    . You can change skin tone and hosiery and color there. The attribute HasDressingRoom tells whether a
    product has this feature or not.
29. 7/10/00 Will the reports be printed in color for the "judges" ?

    Yes. Please tell this when submitting.
30. 7/10/00 Please confirm that the following sentence provided in the instructions is still true:

    "In the test set that you will get for question 1, the above set of page views has a 50% of being clipped in the middle."
    This question is asked because an 80:20 distribution was seen in the training data set.
    Yes.

Please note that if a session is chosen to clip, but its length is 1, it will not be clipped actually because the clip point is always great than or equal to 1.

31. 7/10/00 'Send Email?' attribute. What does it mean? Who sends email to whom and when?

    Send Email is an attribute from Gazelle's registration page where the customer registering can specify if he/she wants to receive email from Gazelle.
32. 7/10/00 'Unknown card?' attribute (Acxiom). How can Gazelle use an unknown card?

    The 'Unknown card' attribute from Acixom indicates if the customer
    has a credit card that is not of a known type (gas card, bank card,
    upscale card). So a true here indicates that the customer has a
    credit card but Acxiom does not know the type. This field has nothing to do with the credit card used by
    the customer to purchase products from Gazelle's web site.
33. 7/10/00 'Home Market Value' attribute (Acxiom). Does this attribute have any relevance to e.g. people that does not own their homes?

    Home market value is the value of their current home. They may or
    may not be home owners. In some cases they may be home owners but
    Acxiom may not know about it.
34. 7/10/00 'Number of Credit Lines' attribute (Acxiom). What exactly is a credit line?

    This corresponds to the number of credit cards the customer has (or that
    Acxiom knows about).
35. 7/11/00 What does the column "**Order Amount Sum Percent Having Discount Range 0**" in q3_agg mean?

    For each customer,
    100 * (sum of order amounts which do not have discount) / (sum of all order amounts).
36. 7/11/00 What does the column "**Order Amount Sum Percent Having Discount Range (5 ... 10]**" in q3_agg mean?

    For each customer,
    100 * (sum of order amounts each of which has a discount ranging from
    $5 to $10 (not including $5 but including $10)) / (sum of all order amounts).
37. 7/11/00 What does the column "**Order item Quantity Sum Percent Having Discount Range (0 ... 5]**" in q3_agg mean?

    For each customer,
    100 * (sum of the numbers of items in the orders with a discount ranging from
    0 to $5 ( 0< ... <= $5)) / (sum of the numbers of items in all the orders).
38. 7/11/00 What does the column "**Order Amount Sum Percent On Tuesday**" in q3_agg mean?

    For each customer,
    100 * (sum of order amounts purchased on Tuesday) / (sum of all order amounts).
39. 7/11/00 What does the column "**Percent of Products Purchased on Thursday**" in q3_agg mean?

    For each customer,
    100 * (sum of the numbers of items purchased on Thursday) / (sum of the numbers of all items purchase).
40. 7/11/00 What does the column "**Order Line Amount Sum Percent Hanes**" in q3_agg mean?

    For each customer,
    100 * (sum of order line amounts of products of "/Products/Legwear/Hanes") /
    (sum of order line amounts of all products).
41. 7/11/00 What does the column "**Order Line Quantity Sum Percentage blue**" in q3_agg mean?

    For each customer,

100 * (sum of order line quantity of products with MyLegsColorRef = blue) /
(sum of order line quantity of all products).

42. 7/11/00 What does the column "**Order Line Quantity Sum Percent With DressingRoom**" in q3_agg mean?

    For each customer,
    100 * (sum of order line quantity of products with HasDressingRoom = true) /
    (sum of order line quantity of all products).
43. 7/11/00 What does the column "**Order Line Quantity Sum Percent Of Childrens Dance Collection**" in q3_agg mean?

    For each customer,
    100 * (sum of order line quantity of products belonging to Children's Dance Collection) /
    (sum of order line quantity of all products).
44. 7/11/00 What does the column "**Order line Day of Week First:**" in q3_agg mean?

    For each customer,
    the day of week when the first order line is submitted.
45. 7/11/00 In Question 1 aggregated data, why all the following columns have value 0 in some rows?

    Num main Template Views, Num products Template Views, Num articles Template Views, Num account Template Views, Num checkout Template Views. It seems that there are only these 5 template categories.

    These columns were generated through pivoting the number of page views by the "Content Level 2 Path" column. This column has 7 possible values: main, products, articles, checkout, account, Replenish, and include. Due to that Replenish and include occur in very small percentages of records in the training set (0.03% and <0.01%), we didn't generate new columns for them. Besides, the "Content Level 2 Path" column may have null value if errors occurred and no template was created during page view.
    Therefore, it is possible that in some rows all of these columns have value 0.

46. 7/13/00 For question 2, is there a difference between predicting 'Other' (The user will visit another product detail page.) and predicting 'NULL' (The user will not visit any product details page at all.)? The 'NULL' prediction incorporates that the visitor only visits other pages than product detail pages in the rest of his session or that the visitor does not visit any further pages at all, i.e. the given sessions ends.

    The prediction for each session must be one of the 4 classes: "Hanes", "DonnaKaran", "AmericanEssentials", and "Other". "NULL" mentioned in the class should be included in "Other".
    Question 2 has 4 classes:
    - **Hanes** (View Hanes Later)
    - **DonnaKaran** (View DonnaKaran Later)
    - **AmericanEssentials** (View AmericanEssentials Later)
    - **Other** (including any situation that does NOT have Hanes brand views, nor DonnaKaran brand views, nor AmericanEssentials brand views later. This includes no further page views.)

The prediction for each session must be one of the 4 classes: "**Hanes**", "**DonnaKaran**", "**AmericanEssentials**", and "**Other**". "NULL" mentioned above should be included in "Other".

47. 7/13/00 Why is Order Amount not equal to the sum of Order Line Amount for a given order? When the orders dataset is rolled up by order, the mean of Order Amount is 14.0184224, yet the sum of Order Line Amount is 20.8958044.

    Two factors here:
    - Order line amount does not include shipping charges. Shipping is assigned to an entire order, but is not distributed to order lines.
    - Discounts can be applied at the order level or the order line level. Gazelle had a lot of sales with $10 off order level discounts. This probably accounts for the average sum order line amount being higher than average order amount.
48. 7/14/00 What does the column "**Percent Order Promotion FRIEND**" in q3_agg mean?

    For each customer,
    100 * (the number of orders with the "Friend" promotion) / (the number of all orders).
49. 7/14/00 What does the column "**Order Date Recency**" in q3_agg mean?

    Days since last order.
50. 7/14/00 What does the column "**Order Date Frequency**" in q3_agg mean?

    Expected number of orders per year (estimated based on orders so far).
51. 7/14/00 What does the column "**Height Sum**" in q3_agg mean?

    The sum of heights of products viewed by this customer. Note that some products may not have the Height attribute.
52. 7/14/00 What does the column "**Collection Last**" in q1_agg mean?

    The value of the attribute "Collection" of the product viewed in the last page if any. Some example values of the Collection" attribute are: Specialty Items, Childrens Dance, and Oroblu Fashion Line.
53. 7/14/00 For question 3, 4, or 5, can we write a report of more than one page (say 2-3) that does not exceed 1000 words overall?

    Yes.
54. 7/15/00 What does "**Session visit count**" mean in q1_agg data?

    Number of visits from the cookie at the time of the session.
55. 7/15/00 What does "**Num articles/dpt_about_mgmtteam Template Views**" mean in q1_agg data?

    It's the number of pages viewed with "Content Level 3 Path" = "/Content/templates/articles/dpt_ablut_mgmtteam.jhtml".
56. 7/15/00 What does "**Num main/login2 Template Views**" mean in q1_agg data?

    The number of pages viewed with "Content Level 3 Path" = "/Content/templates/main/login2.jhtml".
57. 7/15/00 What does "**Num main/registration Template Views**" mean in q1_agg data?

    The number of pages viewed with "Content Level 3 Path" = "/Content/templates/main/registration.jhtml".
58. 7/15/00 What is the difference between assortment views and product views?

    A product view tells a page viewed containing the product.
    A assortment view tells a page viewed containing a product of that assortment or a page related to that assortment. In addition, the assortment id in the clickstream data is sticky in the sense that, once it is set for viewing an assortment, it remains set on subsequent page requests until a new assortment is viewed. So, assortment id on a page request actually indicates the latest assortment viewed in the session.
59. 7/15/00 What does "**Session Last Template Top 5**" mean in q1_agg data?

    The request template of the last page viewed (in the dataset rather than in the true session). This column has the template if it is one of the 5 most common values; "Other" otherwise.

60. 7/15/00 What does "**Session First Referrer Top 5**" mean in q1_agg?

    The referrer of the first page of the session. This column has the referrer if it is one of the 5 most common values; "Other" otherwise.
61. 7/17/00 Why does the average first prossessing time is 752 ms in the question1 aggregated data set and only 292 ms in the question1 aggregated test set?
    Gazelle changed their site, and it's noticed that the compile time is reduced. In addition, the site simply became more stable. Every time the application server is restarted, or when new content is staged, the first access to a page causes a recompile of that page's jhtml, which takes about 5-10 seconds. If you look at access times for first pages excluding pages that took more than 5 seconds, you'll see

    training set: 153 ms
    test set: 130 ms
    much closer.For up to 10 seconds, it's 208 ms versus 161 ms. Again, reasonably close.
62. 7/17/00 There are a few new content templates in the test data set:
    - Content/templates/account/create_credit.jhtml
    - Content/templates/account/edit_credit.jhtml

    Do these templates replace other templates?
    If yes: Can You please name the previous templates' names.
    If no: What is the busines function of these new templates?
    On the other hand, there are a few content templates missing in the test set that were present in the training data set:
    - Content/templates/account/billing_info.jhtml
    - Content/templates/checkout/billing.jhtml
    - Content/templates/checkout/creditcard.jhtml
    - Content/templates/checkout/shipping.jhtml
    - Content/templates/main/update_credit_card.jhtml
    - Content/templates/main/update_user_address.jhtml

    Did You replace these templates by templates that provide the same functionality but that have a different name?

    The end client, i.e., Gazelle, is doing these changes, not Blue Martini. Blue Martini provides the infrastructure to build the web sites, but the clients build the .jhtml and change them often. This is one of the problems with a live web site and the main reason that Blue Martini logs events and clickstreams at the semantic level, including products, assortments, etc.

    Gazelle streamlined some of their processes between the train and test period, which caused some of their templates to change.

    - account/create_credit.jhtml Used for creating a credit card.
    - account/edit_credit.jhtml Used for updating a credit card.
    - account/create_address.jhtml Used for creating an address.
    - account/edit_address.jhtml Used for updating an address.

    Most of their checkout procedure has been replaced by a streamlined express checkout which uses the checkout/expresscheckout.jhtml template.
63. 7/17/00 What does "**Session First Request Date**" means?

    Generally, if the string has First or Last then it is the first/last entry from the unaggregated data. This is therefore the date+time of the first request in the session.
64. 7/17/00 What does "**Collection Last**" means?
    Collection from the last record in the session. Collection is a Gazelle attribute, such as Women's Summer Spectator Collection.
65. 7/17/00 What does "**Session First Content ID**" means?

    Content ID is a unique identifier for the template used to display the page.
66. 7/17/00 What does "**Num BrandOrder Assortnment views**" means?

    The column assortment path in the unaggregated data can take multiple values. In the aggregated data,

this is a pivot/count of the value BrandOrder, i.e., we count how many times BrandOrder appeared in the assortment.

67. 7/17/00 Does the event of purchasing occur at the ExpressCheckout page? or at the confirm_order page? In other words, when is the shopping information sent to the site?

    At the confirm_order page.