 **Data Mining Cup**

**Case: Winter 2015**

**Background**

Beyond housing an impressive collection of piano playing cat videos, the internet is the established go-to consumer research platform. Paid internet search (dominated by giants such as Google, Yahoo!, and Bing) is a significant marketing tool that provides varying methods for businesses to interact with consumer search activity.

In this competition, you will use your creativity and statistical problem solving skills to develop a solution that will evaluate the interaction between paid search impressions and the return on marketing investment (profit). We will ask you to leverage insights about keywords and bids to build an optimal strategy (a system of models and/or business rules) for identifying the best keywords by predicting their breakeven bidding amounts.

You are a member of an analyst team within StatsBank tasked with improving the effectiveness of Search Engine Marketing (SEM). You and your team members will be provided data that can be used to create a solution which will optimize SEM investments. You are encouraged to exercise creativity in data transformations, data aggregation and innovative modeling techniques to solve the problem.

Suggested Project Plan

As this case has a large volume of data available for use and a component of variable creation, below is a suggested project plan to guide the amount of investment allotted to each step in the strategy development process.

**Step 1 – Strategy Design**

* Review case details and determine the conceptual framework for your strategy
* Outline how your approach solves the business problem

**Step 2 - Data Exploration & Variable Creation**

* Understand the properties of the sample data and its assumptions / limitations
* Determine methods of data handling, aggregation & transformation
* Create the variables and features needed to support your strategy
* Collect alternative data sources to compliment the supplied data
* Purge un-useful and redundant data, if there’s any.

**Step 3 - Variable Selection & Strategy Development**

* Determine the modelling technique your group wants to pursue.
* Finalize the inputs and specifications
* We recommend you iterate a bit so you can incorporate learnings

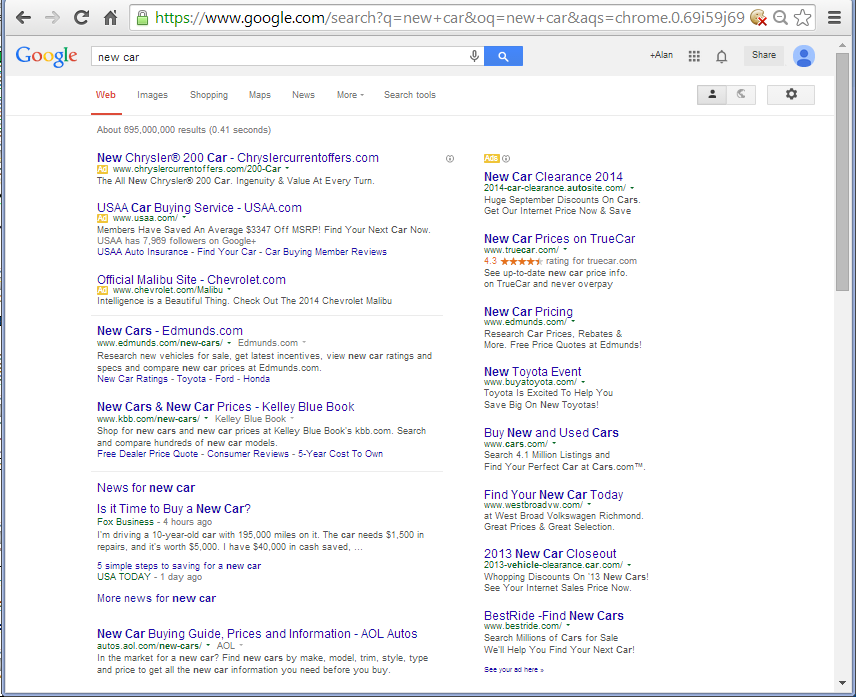
**Step 4 - Strategy Assessment & Presentation Preparation**

* Analyze model performance and produce breakeven biddings at keyword level.
* Prepare a presentation that describes the steps your group took and the business insight based on the result.

If you have any questions on the case materials provided, the data or your analysis, please contact the Capital One mentors.

**Paid Search Mechanics 101**

SEM allows businesses to relate their product offerings to search terms and compete for one of the eight to twelve paid placements that are prominently presented in the search engine results page.



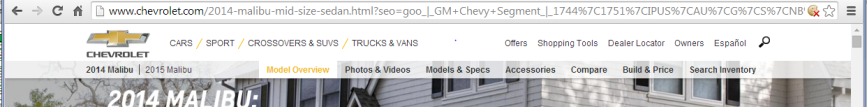
**Search Query**

**SEM Advertisements**

**Organic Search Results**

**External ID**

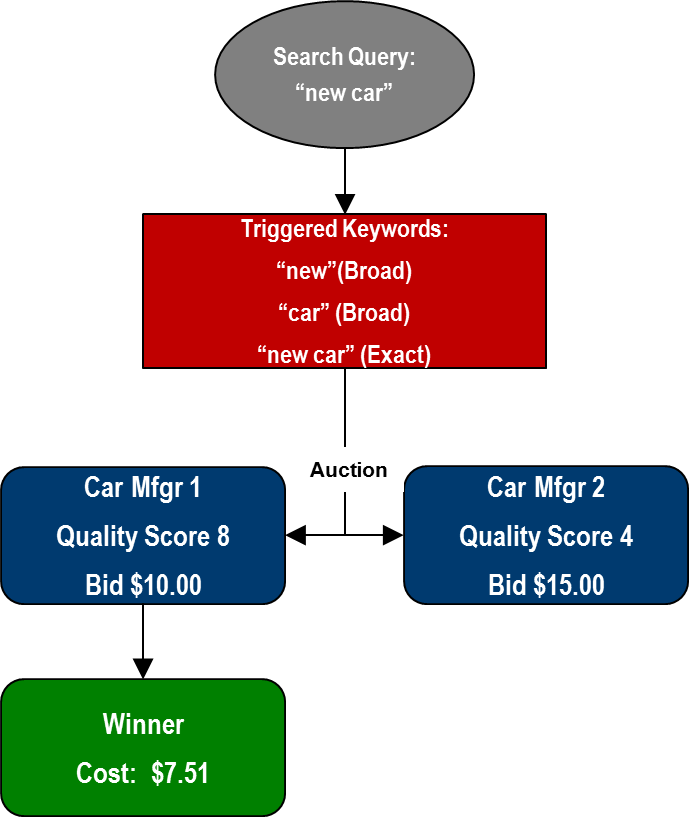
**Click**



The linking of search terms, or keywords, and the paid placement is facilitated by an auction mechanism hosted by the search platform. StatsBank pays the search platform on a per-click basis when StatsBank’s ad is clicked by the consumer conducting the search. Further, the auction process is affected by “quality scores” which are calculated by each search platform (using a proprietary algorithm) to quantify the relevance of the ad content.

The bid amount interacts with the ad placement evaluation and price per click derivation according to a process dictated by the search engine platform:

1. Consumer searches for “new car”
2. Keywords are tokenized and evaluated
3. The Bid Amount is normalized by the Keyword-Quality Score and compared to other bids
4. Car Manufacturer 1 wins the bid (see diagram below), and gets the top spot on the search results
5. Should the ad be clicked (paid advertising only generates marketing expense on a per clicked ad basis) the amount per click is calculated via a variant of a Vickrey Auction method: $0.01 more than the 2nd place bid amount adjusted by a ratio of the Quality Scores



**Bids & Costs are normalized by Quality Score:**

**Bid \* Quality Score**

**Car Mfgr 1 Car Mfgr 2**

**=$10.00 \* 8 =$15.00\*4**

**=$80.00 =$60.00**

**$80.00 > $60.00 → Car Mfgr 1 Wins**

**Car Mfgr 1 Cost per Click**

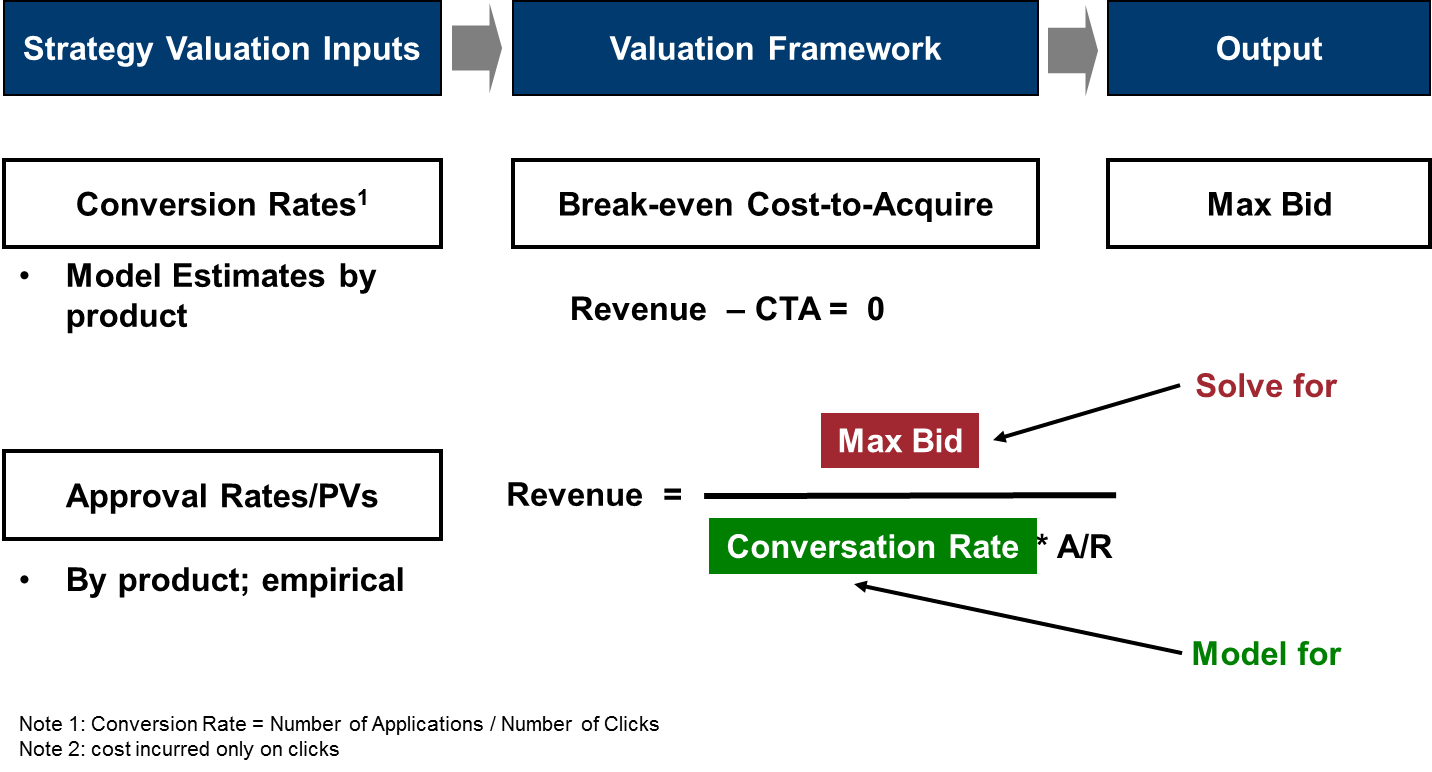
**BidCarMfgr2 \* (QSCarMfgr2/QSCarMfgr1) + $0.01**

**=$15.00 \* (4/8) + $0.01**

**=$7.51**

Additional details of the auction process are well documented at search platform hosts (google.com, yahoo.com, bing.com). For example: Google offers a mini-course highlighting their search engine marketing product (called AdWords): https://www.google.com/onlinechallenge/dmc/dmc-seo.html

**Breakeven Bids Explanation**



**Scenario – You want your Bid Strategy to Break Even; determine the Maximum Bid**

Revenue per Account = $150

Approval Rate (A/R) for Product = 60%

For Keyword String “Word1 +Word2” Conversion Model predicts 10% of Clicks result in applications for Product

$150 \* ( 10% \* 60% ) = $9.00 -- Your maximum bid per click would be no more than $9.00

For Keyword String of “Word3 +Word4” Conversion Model predicts 2% of Clicks result in applications for Product

$150 \* ( 2% \* 60% ) = $1.80 -- Your maximum bid per Click would be no more than $1.80

You will realize that there’s no approval rate and revenue information in the validation dataset as your team go through the dataset. This is intended to be a challenge for your team. Be sure to attend to the Q&A session if you have questions about it!

**Problem Formulation**

Your VP is under pressure to ensure that marketing investments are performing as efficiently as possible. She knows that StatsBank participates in the paid search marketplace to prospect for customers, as do competitors of StatsBank. Presently, StatsBank leverages an empirical solution rather than a statistically optimized framework, but the VP is now asking you to create a more rigorous statistically based solution. Our in-house talent is certainly capable of developing an improved solution.

Your VP points out that in order to build an effective strategy, it is essential that we are able to evaluate the following:

* Relationships between Keywords and Ad Groups
* Keyword bid amounts
* Cost & quality of clicks

… and of course, it is imperative to ensure that our Non-Technical business partners understand our solution, how it adds value, and its limitations.

With the analytic tools you create, your team is being asked to determine the breakeven bidding amount for each of the keyword we are currently bidding on.

**Statistical Package and Methods**

You can apply any statistical software package (SAS, SPSS, R, etc) for the modeling portion of the competition. Any statistical method can be used to build your model, such as logistic regression, linear regression, decision trees, SVM, neural network, etc…

In addition, you are told that one requirement for your work is to understand the features that are used to predict the breakeven bidding amount (why did you select those variables and why do you think they are relevant?)

**Data Access**

Data will be made available for download online at <http://capitalonecampus.ca/>.

**Case Q&A**

There will be a Q&A session hosted on Saturday starting at 10 am where we will go through the case in detail and answer questions your team may have. In order to keep things fair and square, questions will not be answered via email.

**To Submit Results**

Each team (with a maximum of four members) needs to submit three files for a qualifying entry: your dataset, your code, and your slideshow presentation

1. The **dataset** should:
   1. be in a simple file format (.csv, .dat, .txt, etc...)
   2. contain the following columns (and all associated observations) from the validation dataset (SEM\_DAILY\_VALIDATION.csv):

WEB\_ACTVY\_DT            - Date of Record

ENGN\_ID                       - Indicator of Search Engine (e.g. Google, Yahoo/Bing)

LANG\_ID                       - Language indicator

DVIC\_ID                        - Targeted Device (e.g. Mobile, Table, PC)

CMPGN\_NM                  - Campaign Name

AD\_GRP\_NM                 - Ad Group Name

KEYWD\_TXT                  - Keyword Text String

**CR\_PRED                       - This will be your conversion rate predictions**

**BE\_BID                         - This will be your recommendations for break-even bids**

1. Your **code** should :
   1. run in a common statistical language (SAS, R, C, etc...)
   2. be able to reproduce your results
2. Your **slideshow presentation** should:
   1. be a business proposal that includes a description of your approach, method and strategy
   2. be a maximum 15 slides not including appendix (aim for 10-15 minutes in delivery)
   3. have at least one slide describing your approach and strategy
   4. have at least one slide describing the statistical methods used and the steps in the model-building process
   5. have at least one slide on the interpretation of the statistical, business insights, and long term strategy
   6. have all non-essential material in an Appendix section at the end of the slideshow
   7. be compliant with the contest rules

We will ask each team to submit their entry via email to dataminingcup@capitalone.com by 10:00:00 am, Monday, February 16th, 2015.

**Judging**

There are three rounds of judging.

*Round One*

All entries received by the submission date will be reviewed by a panel of Capital One associates who will select the top five (5) teams from each participating location to move onto the next round. Decisions will be made based on the following criteria:

* Performance of the proposed strategy, evaluated based on the accuracy of predicted breakeven bidding amount per keyword
* Articulation of strategy in slideshow
* Evidence of analytical rigor and creativity

The selected teams will be invited to the semi-final round of judging. Teams will be notified of Round One judging results by 4:00pm, Monday, February 16th, 2015.

*Semi-Final*

The five (5) semi-finalist teams will be invited to give a presentation on their business strategy proposal to a panel of leaders from Capital One who will be on-campus Tuesday, February 17th, 2015. Presentations will be judged on the following criteria:

* Clarity and organization of thought
* Overall presentation skills
* Demonstrated analytical ability

The top 2 teams from each participating locations will be selected by the judging panel based on the overall persuasiveness of the proposal. These teams will be invited to attend the final round.

*The Final*

The invited team from each of the locations will be invited to give their presentation to the senior leadership of Capital One Canada at the Capital One Canada office in North York, Toronto on Wednesday, February 18th, 2015. The judging criteria remains largely unchanged from semi-final round, but be prepared for more tough questions!