SFU Beedie Business Analytics

HACKATHON

SFU BEEDIE SCHOOL OF BUSINESS
BADM
Business • Analytics • Decision • Maker

Generously supported by:
VESTA PROPERTIES

SFU Beedie Business Analytics Hackathon 2019

1. Case Context

Liberty Mobile (LM)[1] is one of the Canadian wireless telecommunication service providers. Unlike the big three (i.e. Rogers, Telus and Bell) and their subsidiaries, LM does not yet have national coverage and currently covers only some municipalities in two western provinces, BC and Alberta. In addition to lower-price phone plans, LM competes with the big three on a unique selling point, namely Data at Liberty (DAL): all LM customers pay no overage charges. When a customer exceeds the data limit of a plan, the customer would just experience throttled data speed.

DAL was introduced two years ago and has been a huge success, allowing LM to build up its customer base. Unfortunately, in June this year, the big three started offering unlimited data plans, which are essentially DAL (no overage charges and only throttled data speed beyond the data limit of a plan). These unlimited data plans have become very popular, essentially neutralizing LM's once unique DAL. LM has since lost 12% of their customers. As the general rule of thumb in this industry is that it is 10 times more expensive to get a new customer than to keep one, the top management at LM sets it a top priority to stop the bleeding. VP Marketing is thus tasked with coming up initiatives to stop the customer churns. The first step to it is to understand which customers are most likely to churn, and identify customer characteristics that are most associated with the churns. The insights will then be used to develop retention, upselling or cross-selling initiatives.

While you are free to conduct any analysis with the below data variables, each team is expected to build a statistical/machine learning model, which can predict the churns in a random sample of LM's post pay customers in last two months (Sep & Oct).

2. Datasets

There are three datasets:

1) ACCOUNT_DATASET: 4000 accounts are sampled; Note that an account may have multiple phone lines (because a customer can get $50 one-time rebate each time s/he adds a new phone line to an account)

[1] Although the company is fictional, the case context is based on real industry events and situation. In other words, please feel free to research about the Canadian telecom industry like unlimited data plans, if you are not familiar with it.

2) PHLINE_DATASET: The usage data of all the phone lines from the 4000 accounts are extracted
3) CITY_DATASET: The average usage data of LM customers at different cities and towns in BC and Alberta

Variable	**Description**
acc_num	Account number
st_date	Starting date of the account
cust_age	Age of the account's billing customer
bill_city	Billing address (city) of the account's billing customer
cr_score	Credit score of the account's billing customer
ph_num	Phone number
ph_k_date	Starting date of the most recent phone plan contract, associated with the phone number
mon_data	Monthly data consumed (GB) in Jan-August 2019 by the phone number
data_plan_m8	Data plan in August 2019, subscribed by the phone number (all data plans include unlimited voice across Canada and unlimited SMS)
disc_m8	Monthly discount in August 2019, assigned to the phone number
mon_sms	Monthly number of sms sent in Jan-August 2019 by the phone number
mon_voice	Monthly number of minutes in voice in Jan-August 2019 by the phone number
serv_tick_m1to6	Monthly number of service call tickets in Jan to June 2019 by the phone number
serv_tick_m7to8	Monthly number of service call tickets in July & August 2019 by the phone number
data_roam	Annual number of days for data roaming in the U.S. and internationally ($7 per day) by the phone number
long_d_min	Monthly number of minutes in long-distance calls in Jan-August 2019 by the phone number
long_d_spend	Monthly spending in long-distance calls in Jan-August 2019 by the phone number
tot_pay	Monthly total spending in Jan-August 2019 by the phone number
churn	= 1 if the phone line of the phone number is terminated in Sep & Oct 2019 = 0 otherwise
sample	Estimation, Validation, or Holdout
mon_data_city	Monthly data consumed (GB) per LM phone line in the city/town in Jan to June 2019
mon_sms_city	Monthly number of sms sent per LM phone line in the city/town in Jan to June 2019
mon_voice_city	Monthly number of minutes in voice per LM phone line in the city/town in Jan to June 2019

serv_tick_m1to6 _city	Monthly number of service call tickets per LM phone line in the city/town in Jan to June 2019
serv_tick_m7to8 _city	Monthly number of service call tickets per LM phone line in the city/town in July & August 2019
data_roam_city	Annual number of days for data roaming in the U.S. and internationally per LM phone line in the city/town
long_d_min_city	Monthly number of minutes in long-distance calls per LM phone line in the city/town in Jan to June 2019
long_d_spend _city	Monthly spending in long-distance calls per LM phone line in the city/town in Jan to June 2019
tot_pay_city	Monthly total spending per LM phone line in the city/town in Jan to June 2019

Notes:

- You're free to merge the three datasets and construct any variables you believe to be helpful
- There is no oversampling or undersampling of any of the datasets
- Any missing value will be identified as "NA". You're free to handle the missing values in any way you see fit
- 25% of the values in the target variable, *Churn* is "NA" as these 25% are the holdout sample for the predictive model assessment (Please see the below Technical Competence Score section). In other words, you observe the *Churn* values for 75% of your observations and need to predict the *Churn* values in the remaining 25%. A variable named *Sample* is included to easily distinguish each case. For the holdout sample, the *Sample* variable value will be "Holdout". This will allow you to perform any data cleaning and manipulation on both holdout and non-holdout (your training) data simultaneously without needing to have separate files for each, so that the models you build will be readily applicable to the holdout sample. *To reiterate, any records that have Sample labeled as "Holdout" have the target variable missing.*
- For the non-holdout data, it is further divided randomly into "Estimation" and "Validation" with a 2/3 vs. 1/3 split. This is done simply to facilitate your own modeling on a subset of the data (the Estimation set) and testing how well it does with a known target on another subset (the Validation set). However, you may ignore this distinction and use the data with the known target however you like.
- You will add a new variable to the data labeled "score", which has the predicted probabilities for *Churn* for each row, from your best model.
- You may use any software.
- Don't spend the whole time during deliberation (9am to 1pm) building and refining your predictive models. Judges will be looking at criteria outside of technical model building skills (Please see the next section).

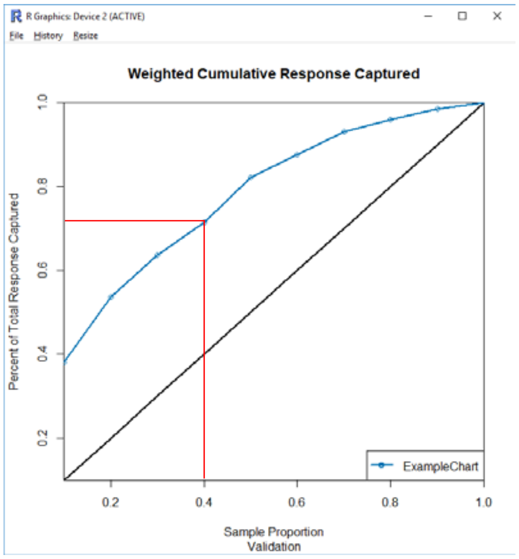
3. Performance Assessment

Each team will be assessed on three components, namely technical competence, business understanding and communication. The technical competence will be assessed by your team's best predictive model performance on the leaderboard.

3.1 Technical Competence Score

Using your best models, you will assign probability scores for the event, *Churn* in the holdout sample (where the true churns are unknown to you). This scored dataset will be submitted to the website https://beediehackathon.bus.sfu.ca/, which will compare your probability scores to the known true churns. Specifically, the probability scores and the true target will be used to generate a cumulative captured lift chart. The final value for comparison will be the percent captured at the 40% sample proportion level.

Below is an example of a "cumulative captured" lift chart, and the 40% captured level is shown here:



This model is capturing about 72% of the positive responses in the top 40% of predicted probabilities.

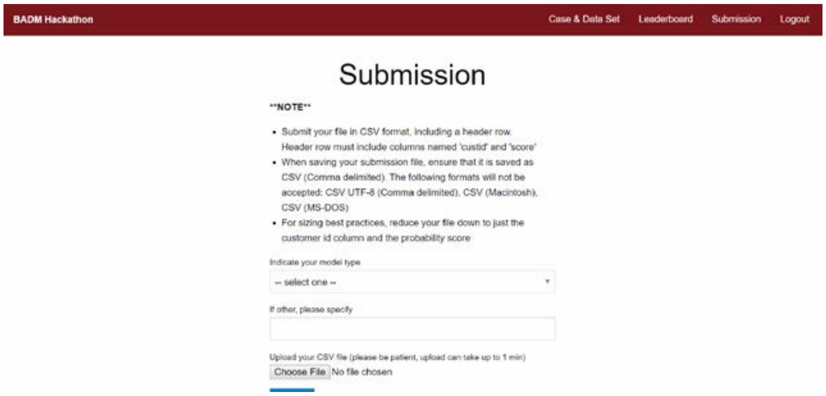
Teams do not need to generate or use lift charts. You may use any model quality measure you are comfortable with (hit rates, AUC, etc.) when building your best models, but all models will be assessed with the above method.

Notes:

- Save the scored dataset as a ".csv" file, with the filename "*your teamname.csv*". e.g, if you are using R,

  write.csv(Dataset, "C:/yourdirectories/…/yourteamname.csv")
- **Ensure that all csv files are NOT csv UTF-8** (e.g., choose the right csv format if using Excel)
- **Ensure the submitted file has only *ph_num* and *score* (all lower case)**

- You should have received your team’s login credentials to https://beediehackathon.bus.sfu.ca/ at registration. After logging in, you can submit a **maximum of three datasets** of predicted probabilities throughout the competition.
- On the website, navigation is done using the three links in the top right corner
    - Case & Data Set – access the data set from here
    - Leaderboard – see how your model is performing
    - Submission – submit a model score
- When submitting, ensure you follow guidelines on the file format and size requirements of the submission:



Submission

NOTE

- Submit your file in CSV format, including a header row. Header row must include columns named 'custid' and 'score'
- When saving your submission file, ensure that it is saved as CSV (Comma delimited). The following formats will not be accepted: CSV UTF-8 (Comma delimited), CSV (Macintosh), CSV (MS-DOS)
- For sizing best practices, reduce your file down to just the customer id column and the probability score

Indicate your model type

-- select one --

If other, please specify

Upload your CSV file (please be patient, upload can take up to 1 min)

No file chosen

- A real time leaderboard will display progress of competitors until noon but you can submit your predictions until 1pm. The top team will be the one capturing most churns at the top 40% of the holdout sample, sorted by their probability scores in descending order.
- You are encouraged to submit your predicted probabilities early, to ensure that you have something that will give you a chance to get into the storytelling portion of the competition, and to continue to try and improve your model predictions.
- The percentage of churns captured by your best model (out of the three submissions) at 40% of the holdout sample will be converted to a technical competence score for your team.

3.2 Business Understanding & Communication Scores

You should also prepare a slide deck before 1pm. The slide deck will then be used to present your churn model and any other additional analyses to several industry judges. The goal is to explain how the analyses can support the decision and/or address the issue in the case. Specifically, the industry judges will play the role of VP Marketing, who **prefers non-technical but yet evidence-based arguments**. Your business understanding and communication will be assessed by the judges.

There are two rounds of presentations to the industry judges. While every team will participate in the preliminary round, only five teams will present in the final round.

3.2.1 Preliminary Round Judging

- At registration, your team is randomly assigned to a group of five to six teams. You will be competing against teams inside your group. Only one team from each group can advance to the final round. There are in total five groups and thus five finalists after the preliminary round.

- You will pitch your analyses to a set of judges assigned to your group one by one. Specifically, you will have five minutes of conversation with each judge. Be prepared that the judge may interrupt you with questions. It is thus important to identify the most important slide you want to show ("money" slide) and show it in the first minute or so.
- After the five-minute presentation/conversation, the judge will score your team's performance on Business Understanding and Communication. Then, the judge will give you some feedback so that you can make changes for the next judge you'll pitch to!
- The judges of your group will then consider all the judges' assessment of your Business Understanding and Communication as well as your Technical Competence score to compare your team to other teams in the group.

3.2.2 Final Round Judging

- The presentation order of the five finalists are determined by random draw.
- All five finalists will wait outside the presentation room and only enter the room when it is their turn to present.
- Each finalist will have 10 minutes to present their analyses and managerial implications. There will then be 5 minutes of Q&A.
- Judges will then rank the five finalists based on their assessment of their Business Understanding, Communication and the Technical Competence score from the leaderboard.

Notes:

- Email your slide deck to badmfest@sfu.ca by 1pm. Name your file and email subject line as [Team Number] Presentation Slides, e.g., "[Team 32] Presentation Slides.pptx"
- The five finalists will use the submitted slide deck in the final round.
- For the preliminary round, choose a slide from the deck to make it your "money" slide (the slide that provide so much insights that your boss/client would feel you earn your paycheck by just looking at that slide) and build your elevator pitch around it. In other words, don't leave the best parts till the end. Judges can and will interrupt during your pitch with questions.

4. Award

There will be two categories of awards. The top three teams in the final round will be the overall winners of the hackathon while the top three teams on the leaderboard alone will receive technical awards. All cash prizes and the event expenses are generously sponsored by Vesta Properties.

5. Summary & Schedule

[9:00am-1:00pm @ Room 1200-1500]: Deliberation (Case and Data Analysis)

- The maximum of three submissions will be received until 1pm but the leaderboard will be closed at noon
- Email presentation slide deck to badmfest@sfu.ca before 1pm (name your file and email subject line as [Team Number] Presentation Slides)

[1:00pm-1:30pm @ Room 1200-1500]: Lunch

[1:30pm-3:00pm @ Room 1200-1500]: Preliminary Round Judging

- 5-minute presentation/conversation with a judge (elevator pitch around a money slide as a judge may ask questions)
- 5-minute feedback from the judge
- Repeat for each of other judges in your group

[3:00pm-4:30pm @ Room 2800]: Final Round Judging

- 10-minute presentation
- 5-minute Q&A