SFU Business Analytics Hackathon 2021

Generously supported by:

1. Case Context

Meta-Visit (MV)[1] is a startup currently providing a virtual care technology platform to about 100 general physicians (GP) in B.C. Their platform allows GP's to attend to their patients via video calls, audio calls, and text messaging in an integrative manner and in compliance with privacy standards like HIPPA, GDPR, and PIPEDA/PHIPA, while automating backend administration like appointment scheduling and medical record management.

With the interests in digital health solutions among venture capitalists and investors at all time high due to the pandemic's acceleration of digital health adoption, the VP of Growth at MV is tasked with developing a plan to build their own direct-to-consumer (D2C) brand, which directly provide virtual care services to the B.C. population.

When a B.C. resident uses a health care or wellness service (regardless it is virtual or in-person), the service can be paid for in three ways, namely (1) through the provincial government's Medical Services Plan (MSP), (2) by the resident's extended health care insurance, and (3) out of the resident's own pocket. All GP's and most specialists in B.C. are fully covered by MSP, which restricts the maximum number of patients a physician can see each day and standardizes the fees for individual services. On the other hand, non-MSP health services, particularly those can be efficiently and effectively delivered via virtual care, include but are not limited to nonreferral mental health counseling or therapy, non-referral dietitian, speech therapy, art and music therapy for Autism Spectrum Disorder and cancer patients. Given the physician shortage in B.C., MV decides to focus on delivering some of the non-MSP services via virtual care, with a plan to expand into wearable health tracking (prescription reminders or fall detection) or at home lab tests in future.

One of the first decisions the VP of Growth needs to make is WHO the new D2C brand should target. The VP thus recruited volunteers from the patients of the GP's using the MV platform, asking them to fill out a questionnaire, and to give consent to use their non-medical platform usage data. The VP believes that a sample of patients with self-reported data via the questionnaire and platform usage, together with some publicly available geodemographic data, would allow some preliminary insights into the targeting decision. Specifically, the first question the VP would like to answer is WHO are the current "power users" of the virtual care with their GP's?

[1] Although the company is fictional, the case context is based on the real industry situation. In other words, please feel free to research about the Canadian digital health market, if you are not familiar with it.

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 power user in the sample of volunteer patients.

2. Datasets

There are three datasets:

1) USAGE_DATASET: The usage data of the MV virtual care platform from Nov 2020 to Oct 2021
2) SURVEY_DATASET: The self-reported questionnaire data by volunteer patients
3) FSA_DATASET: The StatCan geodemographic data at different FSA's (i.e. the first three characters of a postal code) in BC

Variable Name	**Description**
clinic_id	Clinic ID
clinic_fsa	Forward Sortation Area (FSA): the first three characters of a postal code
pt_id	Patient ID
freq	Number of times using virtual care
power_us	Power user or not (defined as freq higher or equal to 11)
income	Self-reported household income of the patient
age	Self-reported age of the patient
edu	Self-reported education attainment of the patient
perc_health	Self-perceived health = Excellent, Very good, Good, Fair, or Poor
perc_weight	Self-perceived weight = Overweight, Underweight, or Just about right
bmi_class	Self-reported BMI class = Overweight, Underweight, or Normal weight
arthritis	Self-reported arthritis = 1 (yes) or 0 (no)
highBP	Self-reported high blood pressure = 1 (yes) or 0 (no)
diabetes	Self-reported diabetes = 1 (yes) or 0 (no)
stroke	Self-reported stroke = 1 (yes) or 0 (no)
heartdise	Self-reported heart diseases = 1 (yes) or 0 (no)
repstrain	Self-reported repetitive strain injuries = Physical exercise, Lifting/carrying, or On the Job
injstatus	Self-reported injury status = Limiting activities, Treated injury, or Treated by Limiting
physactivityindicator	Self-reported physical activity indicator = None, Active, Moderate, or Somewhat
gave_birth_last5	Gave birth in last 5 years = 1 (yes) or 0 (no)
perc_mentalHealth	Self-perceived mental health = Excellet, Very Good, Good, Fair, Poor
perc_lifstress	Self-perceived life stress = None, Not Very, A bit, Quite, or Extreme
perc_workstress	Self-perceived stress at work = None, Not Very, A bit, Quite, Extreme
care_language	language preferred by the patient = English or other
othercare	Having other health care providers = 1 (yes) or 0 (no)
spend_health	Self-reported spending on health care
pop_fsa	Total population of the FSA
median_age_fsa	Median age of the population of the FSA

hhold_fsa	Total number of households of the FSA
median_income_fsa	Median household income of the FSA
hhold_work_health	Population 15 years or over in the occupations in health of the FSA
avg_spend_health	Average total expenditure of a household in health care in the FSA
avg_dcost	Average total direct costs to household for health care in the FSA (e.g., prescribed & non-prescribed medicines and pharmaceutical products, health care supplies, health care services by health care practitioners in the home, hospital care, nursing homes, and other residential care facilities, weight control programs, quit-smoking programs and other medical services, eye care, dental services, and orthodontic and periodontal procedures)
avg_insur_prem	Average total health insurance premiums in the FSA (e.g., premiums for public administered hospital, medical and drug plans, private health insurance plan premiums, premiums for accident and disability insurance, dental plan premiums)
tot_spend_toba_alco	Total expenditure in tobacco products and smokers' supplies in the FSA

Notes:

- You are free to merge the three datasets and construct any variables you believe to be helpful
- There is no under or over-sampling in the three datasets and the true power user rate is 22%
- Any missing value will be identified as "NA". You are free to handle the missing values in any way you see fit
- 25% of the values in the target variable, *power_us* is "NA" as these 25% are the holdout sample for the predictive model assessment (Please see the below leaderboard section). In other words, you observe the *power_us* values for 75% of your observations and need to predict the *power_us* 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 *power_us* for each row, from your best model.
- You may use any software.
- Don't spend the whole time building and refining your predictive models. Judges will be looking at criteria outside of technical model building skills (Please see the next section).

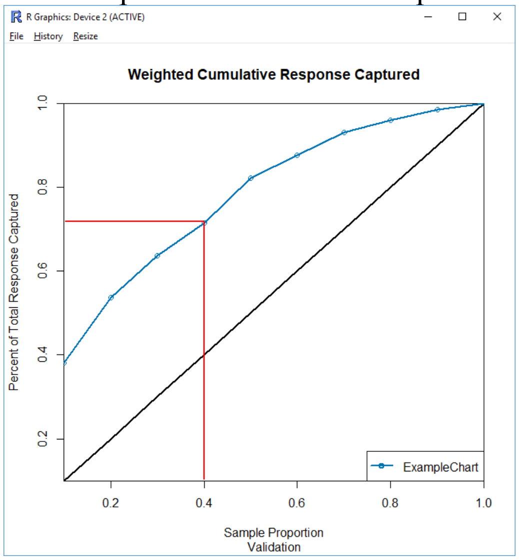
3. Performance Assessment

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 whether a patient is a power user of the virtual care platform in the holdout sample (where the true power user status is 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 values. 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.

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 this method.

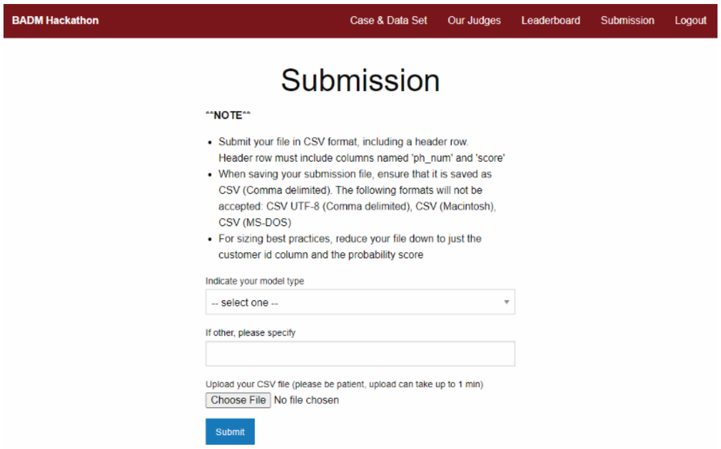
Notes:

- Save the scored dataset as a ".csv" file, with the filename "*your teamname.csv*". For example, 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 *pt_id* and *score* (all lower case)**
- Your team can obtain the login credentials to https://beediehackathon.bus.sfu.ca/ at 9am on Nov 20th from West Mall Centre Atrium. After logging in, you can submit a **maximum of three datasets** of predicted probabilities throughout the competition.

- If there is a tie, the tie breaker will be given to the team submitting the prediction earlier.
- 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 'ph_num' 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)

- 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 power users 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 power users captured by your model at 40% of the holdout sample will be converted to a technical competence score.

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 patient predictive 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 the VP of Growth at MV, who **prefers non-technical but yet evidence-based arguments**. Your business understanding and communication will be assessed by the judges.

Specifically, 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

- Your team will be randomly assigned to a group of four teams when you obtain your leaderboard login credentials. 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 will be determined by random draw.
- All five finalists will wait outside the presentation room and only enter back 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:

- To help industry judges remember your team after your presentation, please take a screenshot of all your team members' faces on Zoom and email it to badmfest@sfu.ca by 10am. Name your file and email subject line as [Team Number] Team Photo, e.g., "[Team 404] Team Photo.jpg"
- 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 404] Presentation Slides.pptx"
- The submitted slide deck is expected to be the SAME as the one you will use in both the preliminary and final round.
- For the preliminary round, you may want to 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.
- Don't leave your breakout room starting from 1pm until you finish talking to all the judges in your group (there should be four cycles of presentations in the preliminary round). Once you finish your preliminary round, please go to the presentation room in WMC.

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 @ WMC Atrium]: Deliberation (Case and Data Analysis)

- Email a team photo (i.e. screenshot on Zoom) to badmfest@sfu.ca before 10am (name your file and email subject line as [Team Number] Team Photo)
- The maximum of three submissions will be received until 1pm but the leaderboard will be closed at noon
- Email presentation to badmfest@sfu.ca before 1pm (name your file and email subject line as [Team Number] Presentation Slides)

[1:00pm-1:30pm @ WMC Atrium: Lunch

[1:30pm-3:00pm @ WMC Atrium & Your team's Zoom breakout room]: 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 @ Presentation Room at WMC]: Final Round Judging

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