# Project Design Writeup and Approval Template

Follow this as a guide to completing the project design writeup. The questions for each section are merely there to suggest what the baseline should cover; be sure to use detail as it will make the project much easier to approach as the class moves on.

### Project Problem and Hypothesis

\* What's the project about? What problem are you solving?

The project tries to identify what factors drive a regular trialist to start paying for hulu, an online streaming platform, within 30 days post initial signup date. The variables being considered include profile-level attributes, acquisition channels and engagement patterns.

\* Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?

This is a classification problem, with the outcome variable being predicted as a binary value. “1” stands for “conversion”. “0” stands for “churn”.

\* What kind of impact do you think it could have?

It’s going to help my company better understand what are the most impactful drivers for promo-to-pay rate.

\* What do you think will have the most impact in predicting the value you are interested in solving for?

Acquisition channel, viewing intensity of content, the variety of content viewed

### Datasets

\* Description of data set available, at the field level (see table)

|  |  |
| --- | --- |
| Userid | Unique user id for each user selected, only users on standard product plan will be selected |
| Gender | Self-reported gender info, may contain null |
| Age | Self-reported age info, may contain random values or null values |
| Acquisition channel | The promotion channel through which the customer was initially acquired |
| Minutes watched per promo day | An indicator of average viewing length during promo period, calculated as total minutes / (first invoice date – signup date) or total minutes / (promo cancel date – signup date) |
| Genres watched per promo day | An indicator of the variety of content watched during promo period, calculated as the number of total genres watched / (first invoice date – signup date) or total minutes / (promo cancel date – signup date) |
| Signup date | When user first signs up as a trialist |
| First Invoice date | The date a promo user converts to a paying customer |
| Canceled\_date | Date when the user canceled the subscription |

\* If from an API, include a sample return (this is usually included in API documentation!) (if doing this in markdown, use the javacription code tag)

### Domain knowledge

\* What experience do you already have around this area?

I work closely with these data at my day-to-day job

\* Does it relate or help inform the project in any way?

I’m very familiar with the caveats of each data fields and how they are captured

\* What other research efforts exist?

\* Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.

\* This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."

\* Include a benchmark, how other models have performed, even if you are unsure what the metric means.

The data science team has done a prediction model using XGBoost earlier this year with the goal of predicting convert-to-pay rate using 3 months of user activity data. Predictions w.r.t. data in first three weeks (i.e., promo. days > 22) are more accurate, although less useful. False Positive Rate (0.6) remains to be high in all the experiments, which may due to the imbalance between samples of converted / unconverted users.

### Project Concerns

\* What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).

- What are the pitfalls of working with self-reported data, in this case age and gender? How should I deal with them?

- If I have 7 types of acquisition channel, will the dummies still work if these end up creating too many new columns, in turn affecting model performance?

- If it necessary to add another variable to bring in top genre watched? My concern is there are too many genres and I’m unsure about how to properly convert them into meaningful classifiers.

\* What are the assumptions and caveats to the problem?

By taking a cross-sectional dataset, I’m assuming the conversion activities are not seasonal based. I’m also assuming the conversion activity is not highly influenced by content availability on the service.

\* What data do you not have access to but wish you had?

- income level

\* What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample (say, predicting who would win an election)

We’re only looking at users who have had some sort of viewing behavior during their promo period.

We’re also including users who are forced to cancel due to payment issues and classify them as promo cancels.

\* What are the risks to the project?

\* What's the cost of your model being wrong? (What's the benefit of your model being right?)

Cost: A waste of spend for the marketing team and wrong strategic prioritization of the company.

Benefit: Help the company meet revenue goal of the year. Improve cost/resource efficiency of marketing initiatives.

\* Is any of the data incorrect? Could it be incorrect?

Since age and gender are incorrect, they would be false.

### Outcomes

\* What do you expect the output to look like?

Co-efficiencies for each predictor

\* What does your target audience expect the output to look like?

They’d expect a ranking of the importance of each variable in terms of its capability to drive the promo-to-pay rate and actionable recommendations.

\* What gain do you expect from your most important feature on its own?

(Not sure if I understand this question;)

\* How complicated does your model have to be?

Mid-level; There is a need to create some binary variables and multi-value columns.

\* How successful does your project have to be in order to be considered a "success"?

AUC being relatively health (is 0.65 a good goal?)

\* What will you do if the project is a bust (this happens! but it shouldn't here)?

- Change to a different model, for example random forest