**03/13 Team Meeting 2:**

Demographics -- Income, age, Gender -> make these columns in themselves.

See how many distinct users there are - is there a prediction for each user (income, age, gender, etc.) or only 1 for each? How many missing values would there be if we joined on user\_id? Same with Psychological,

See if we can beat their model or use their confidence score as a feature to categorize info on the users?

- We are predicted / presenting the holistic view of the customer. We're creating traits. We don't know how they get their confidence score; so lets not use that. User\_id, show they watched, how long they watched, country code, things we can use. Categorizing users; may have multiple outcomes - create data into a bigger table; feed algorithm into user\_id, genres selected, running minutes, ss\_id, with all this, let's predict

Use the confidence score as a filter to get rid of the 'bad' data, pick users that have high confidence scores (say >.5), use that to help us make sure we have good data to go off of.

Going to predict both demographic and psychographic traits. Predicting for level 3 as it implies what level 2 is.

Revenue, turnover rates - business wise. Techinical wise - provide an accurate holistic view by building a classification model to predict for the demographics and psychographics for a given customer. Given an ID, we should know what their demographics and psychographics - what to recommend them, what to send ads for, what to market to them?

\*\* Create Deepnote \*\*

\*\* ask prof to ask more detail on to do what we were \*\*

**03/24 Team Meeting 3:**

Data Prep Questions:

Q: Do we want to bring in part of the demographic data - i.e. - where confidence level is > 99.9%. This could be good data to model off of?

* Something we could do; just take user\_id; trait\_2; trait\_3.
* I can take care of this.

Q: What is the cutoff for confidence score?

* Let’s use .69 - 50th percentile.

Q: Do we get rid of value ‘iflix viewing behavior’ for level\_2? The level\_3 traits that correspond to iflix is player, casual and addict.

* Drop it - doesn’t seem to fit in rest of categories
* Has some kind of bias; all confidence scores of 1.

Q: season\_id; series\_id - what do we do with them?

* Keep them

Q: id's in general - do we convert to strings?

* We can try but not necessary.

Q: Nulls in level 3 -- level\_2 traits are news junkies & avid readers, sports fans, and music lovers -- do we fill level\_3 with these traits?

Q: Nulls in language (701) and genre (5) - get rid of? fill with unknown?

* Get rid of OR match w/ studio ID and asset ID (time allowing\_

Q: do we split dataset up with 1 for movies and 1 for tv shows? (probably not; but maybe we can have different limits for minutes\_viewed?)

* No.

Q: For every record, we, currently, need a level\_3 trait, based off that just one record. This doesn’t take into account what else the viewer may have watched, right?

* KEep it the way it is right now;

Qs: *Feature engineering:*

(1) min minutes viewed - binary column w/ a certain amount of minutes viewed - what is the threshold? 2 min is 25th percentile. 5 minutes makes sense for amount of time to actual ‘watch’ something. Somewhere in between perhaps?

* Keep just one column; 2 min cutoff.

(2) combine genres?

* no.

(3) combine platforms into groups? IE mobile phones; Web Based; Home TV Streaming – Yes, we’ll combine these.

(4) create a threshold for running minutes? – around where the tv and movies cutoff.

—-

Stages:

Prepare Data - df\_psych (demo on standby)

Outliers and missing data.

* ~~Level 3 (specific level 2 traits) – filled level 3 nulls w/ those level 2 traits.~~
* ~~Source language - could use country code to figure out language? - got rid of nulls~~
* Season ID; series ID; Studio ID
  + ~~Delete ones w/o studio ID~~
  + ~~For season ID and series ID - delete~~
* ~~Iflix viewing behaviour as a level\_2 trait~~
* Cut off for confidence score (50th percentile at .69) - after getting rid of iflix view, 50th %ile was at .65. 70th was at .86. Currently, using 70th percentile.

Feature engineering:

~~(1) min minutes viewed - binary column w/ a certain amount of minutes viewed - what is the threshold?~~

* Keep just one column; 2 min cutoff.

~~(2) combine platforms into groups? IE mobile phones; Web Based; Home TV Streaming – Yes, we’ll combine these.~~ This was created using function ‘platform\_type’ to create column of same name.

~~(3) create a threshold for running minutes? – around where the tv and movies cutoff.~~

* Created the threshold for 1 hour. I feel like from a non-technical standpoint, this makes sense. The hour could be the breaking point of deciding to watch something for certain people.

Also - I re-added the demographic data at the end of the notebook where it’s at right now (3/25, 9:30 AM PST) - i created a new df here labeled psych\_demo

Balance Data:

Split Data:

Train Data:

—

Stephen : will do as much as he can tmrw before departure. Will let team know progress.

Dingyi: working code on sagemaker, going to run lab.

Luis: Will work on document and anything leftover Stephen didn’t finish

—-’

3/27/22

Level\_3 traits – either get rid of records w traits with very minimal records

Choose a way to train a model.

1. Use UI (sagemaker autopilot or jumpstart) to load dataset and run the whole training
2. Or train the model step by step.

Currently, getting errors when running a model both ways - errors we don’t really understand.