The project's domain background — the field of research where the project is derived;

Today, more and more businesses are moving to a subscription based model – from cell phone plans to Amazon Prime to Netflix. It's incredibly easy source of recurring revenue, as opposed to a la carte selling. The one big thing that works against subscription based business is customers canceling the service, or churning.

I work at AT&T, and its DirecTV entertainment product is no different. It's a subscription business and we pay very close attention to customer churn.

Here is an example of an academic paper that did something very similar except it is in the fitness membership business:

https://umu.diva-portal.org/smash/get/diva2:1161821/FULLTEXT01.pdf

A problem statement — a problem being investigated for which a solution will be defined;

Since DirecTV is a subscription product, we lose revenue when a customer discontinues service, or churns. By using supervised learning, we can correctly classify if a customer will churn or not. Some of the features that the model will look at will be account level data – customer demographics, package, location, tenure, and October viewership. And of course, if the customer still had service at the end of October or not.

The datasets and inputs — data or inputs being used for the problem;

I will be using a variety of features from the dataset. Some of it is continuous, like the number of hours viewed in a month, number of shows viewed, etc. Other features will be categorical, such as the type of package they have, the gender of the head of household, their income level, etc. There are about 200k customers that voluntarily churn every month. Since there are significantly more accounts that do not churn at the end of the month, I will make sure the volume of churners and non-churners will be balanced. I will make sure I select them randomly.

A solution statement — a the solution proposed for the problem given;

If we can predict whether a customer will likely churn or not, we can proactively use retention offers to reduce churn. I will use supervised learning methods to come up with a model that will classify whether a customer will churn or not.

I will use train_test_split from sklearn.model_selection to split the dataset. Some of the model I am considering right now are SVM, random forest, and naïve bayes.

I may not use the top 95% and bottom 5% of the data to remove outliers. I will use panda functions to normalize data, one hot encode columns, and either remove NAs or turn them into zeros.

A benchmark model — some simple or historical model or result to compare the defined solution to;

The paper that I provided above has an ROC score of around 80%. I would at least achieve that.

A set of evaluation metrics — functional representations for how the solution can be measured;

F1 score would be the primary evaluation metric of the project.

An outline of the project design — how the solution will be developed and results obtained.

1. Create a table that has all the features I need.
2. Clean the data. Normalize and one hot encode columns if necessary
3. Try supervised learning algorithm (SVM, random forest, etc.)
4. Check the F1 score
5. Adjust hyperparameters to improve F1 score