Project Design Write Up – Daniel Walker

**Project Problem and Hypothesis**

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

The problem is related to a client that is developing an all you can eat content consumption app. This app will contain thousands of pieces of content that the user can choose to view. This content is copyrighted material that has been acquired from many different publishers, including several news sources. The content sourced from news publishers will be refreshed regularly to keep current. The content is in the form of both short-form and long-form text.

The client needs to determine which pieces of content should be merchandized to who and at what time. The goal is to maximize each content consumption session with the end customer. The criteria for session maximization includes both total engagement time in the piece of content as well as how far through the content the customer gets (i.e. completion percent).

This is a machine learning problem as we want to use data already gathered around consumption habits and content metadata to predict which future pieces of content are likely to have the best engagement profile. If we are able to reliably predict engagement profiles for new pieces of content before they are live in the app, the content merchandizing team can use this information in their programming decisions. Their programming decisions affect how likely a piece of content is to be discovered by placing them in more prominent positions within the app. So, in essence, we can promote content we feel is most likely to be enjoyed by the customer, creating a better experience and increased likelihood of further customer engagement and retention.

I hypothesize that the ease of reading the piece of content will be one of the major factors that contribute to completion percent. The ease of reading can be determined by running a natural language processing algorithm over the piece of content to extract things like average word length and lexical density.

**Datasets**

The data available is extensive and includes actual customer behavioral data on content current in the app, the text itself, and content and customer metadata. Examples are as follows:

Behavior data:

* Average time in content
* Average completion percent of content

Text data:

* Full text, including titles, copyrights, and chapters

Content metadata:

* Publisher
* Series
* Title
* Genre classification
* Length of content

Customer metadata:

* Reading ability
* Platform (Android vs iOS)
* Days active range
* Free trial customer or paid customer

**Domain knowledge**

I have been working on this project for several months. I understand the data well: how it is generated, how it is stored, how it is access, and how it is defined. I am in the process of mining the data for several different initiatives to better understand customer behavior and how to increase engagement and retention. In addition, there is much public research available around predicting engagement profiles of content.

Current content engagement profiles are not considered at all when deciding which content to promote. Instead, the content merchandizing team group’s content by genre then uses personal taste to decide what to promote. They recognize this is not a data-driven strategy and would be very excited to receive insights about what they should be promoting. The benchmark would be the current average engagement profiles of our merchandized content.

**Project Concerns**

I am not sure how many models to build. It makes sense intuitively that there should be different models to predict engagement for different types of customers. Some customers will have very low reading ability, some will have higher ability, and each will likely need different content to maximize their engagement. So it seems likely I will need to subdivide reading ability and model them separately. In addition, there could be different customer segments that cut across reading ability. For instance, customers that want to learn something when they read vs customers that purely want to be entertained. Additionally, it is clear that completion percent will relate to how long the content is. It does not make sense to try to build completion likelihood of a novel from the completion likelihood of a short news article. Therefore, I will need to group the content into different buckets based on length. In addition, I am not sure how much the data needs to be cleaned. A certain amount of the data needs to be thrown out because there was not enough engagement to provide a reliable view. The next paragraph explains this caveat.

The assumptions with this project are that there are features of the text and metadata that directly relate to how engaged customers are with the content. A caveat to the problem is that currently the product has a relatively low number of users. Some content has never been read (roughly 50% of the catalog) and a lot of the content has low numbers of customers that have engaged. Therefore, we will have to limit our analysis to the subset of content that has had enough users so that an average engagement profile can be established without being overly susceptible to the noise generated from low amounts of data. This means that likely not all genres and all publishers will have enough data to factor into the model. Eventually, more customers will alleviate this problem, however the necessary amount of data take months to acquire. In addition, another concern is that, even with enough customers, the content merchandizing strategy downplays content if it is old or not favorable to the merchandizing team. This could result in some content never being discovered and therefore never having engagement which means it will not contribute to the model.

What is already implied is that completion percent directly relates to the length of the content. I will need to group the content into a few different bucket sizes. More of the catalog is shorter vs longer, so the buckets may need to be over shorter ranges for short-form content, while longer-form content has longer range buckets.

The risks are that it turns out there is not enough content with quality data to be able to build a model. In addition, there is a risk that the customer behavior is more random than anticipated which would make building any model off their behavior less reliable. If the model is wrong, then we will merchandize content that is not any more engaging than a random distribution. It will not improve the customer experience and the current baseline of customer retention will not be improved. There is a low risk of the data itself being incorrect. The instrumentation and metadata recording systems have been thoroughly vetted.

**Outcomes**

I expect the easiest model to build will be a linear regression model that predicts the completion percent of a given title. It will include some publishers, some genres, and some features of the text itself. A lot of publishers, genres, and text features will not have a significant impact on the predicted completion percent. I would like to try other continuous prediction models as well.

The target audience would like a list of content keys in one column and then next to them are the predicted completion percent values. They would rank these values, then classify the content into different genres and promotional sections, then begin their merchandizing strategy at the top of the list.

Do to the complexity of text and the wide divergence of taste and personal preference, I do not expect any one feature to generate a ton of information gain (perhaps less than 10%). The model can have dozens of features, given the richness of the data, however I expect many of the text based features will be highly correlated. This may present a natural limitation on the number of text based features I include.

This project needs to simply increase (statistically significant) the average engagement profile of the app in order to be successful. The more it can be increased, the more successful the project. If the project is a bust, then we will have to consider why. Perhaps the content is less important itself in determining engagement and it is more about the customer. Maybe different types of customers are predisposed to engagement for longer or shorter and you have little ability to influence this. If this is the case, then we need to consider what other features beyond text selection can help influence positive engagement and retention growth.