AI Product: Physical Therapy Exercise Recommendation Application

Capstone Project Proposal

Ben Jacobson
April 2020

Business Goals

Project Overview and Goal

In the physical therapy industry, we will use ML/AI to create recommendations for patient exercise routines, based on physical therapist notes. AI/ML is being used for this task due to (1) data availability, (2) database of labeled physical therapy exercises, (3) available expert knowledge, and (4) the ability to improve physical therapy. This problem has a lot of noisy input data, with many possible exercise recommendations.

The AI application can learn from past expert data, and save valuable therapist/patient interaction time, and lead to a better therapist (and patient) experience, as well as better outpatient products. This is not a diagnostic tool, but an aid to improve the process that generates exercise routines through AI-generated recommendations.

Business Case

There are almost 40,000 clinics throughout the United States, with the 50 largest competitors having less than 25% of the market capitalization. This is a highly fragmented industry and represents an opportunity to leverage AI to improve processes throughout and market AI data products. Considering the possible process improvement through such a product, the business outcomes we seek are:

- Improved physical therapist experience
- Reduced time physical therapist builds exercise plans

As a result of improving the above business outcomes, the following will also be positive outcomes:

- Additional recommendations to select from, offering better care options
- Revenue increase (based on a subscription-based SaaS offering) for our company

Application of ML/AI

The AI model will receive the physical therapist notes as input data (text). Based on this information, the model will predict exercises that should be given to the patient (up to 10 recommended, with ranking).

This is a multilabel classification problem, where multiple exercises may be recommended for a particular problem or symptom. As a result, this should aid the physical therapist is selecting exercises to provide for the patient and save valuable time.

Success Metrics

Success Metrics

The business success metric will be the following:

- An improved physical therapist experience as measured by the Net Promoter Score

As shown in the interface design, the NPS and feedback requests will be built into the user experience.

A baseline metric value will be established initially by surveying physical therapists on their existing product use (whether our current product or not), and gaining NPS score as a baseline. This will also be monitored during testing with stakeholders, and during the initial MVP rollout to a limited number of clinics.

Data

Data Acquisition

The data is sourced from the existing PT Software, which houses both the physical therapist notes, and the recommended exercises. This information, though needing data wrangling, is available in the supporting database. Third-party data will be acquired depending on market need or developed in house (this cost may be substantial).

The initial size of the data is *estimated* at ~20,000 records, of which each record includes the physical therapists' notes, as well as the selected physical therapy exercises. The 20,000 is an estimate of the expected available data based on PT software (existing product). This is a small amount of data for such a task due to the number of potential

afflictions. Also, less likely afflictions will be imbalanced which will need to be remedied in modeling.

This is healthcare information, so potential personally identifying information (PII) concerns must be considered. The product will comply with applicable regulatory requirements (HIPAA), and remove any identifiable data from input data and be applied on aggregate.

Data will become available on a continual basis due to the application's design, so the data science lifecycle will be maintained, and AI product performance monitored and improve over time. The data will be reviewed quarterly, with model development and A/B testing for better AI models. However, during the initial launch, more iterative design changes will occur.

Data Source

Biases built into the data include the diagnosis and recommended exercises that have occurred in the past. For instance, what school of medicine are most physical therapists from? Therefore, it is necessary to have data from numerous, qualified physical therapist clinics to limit this bias.

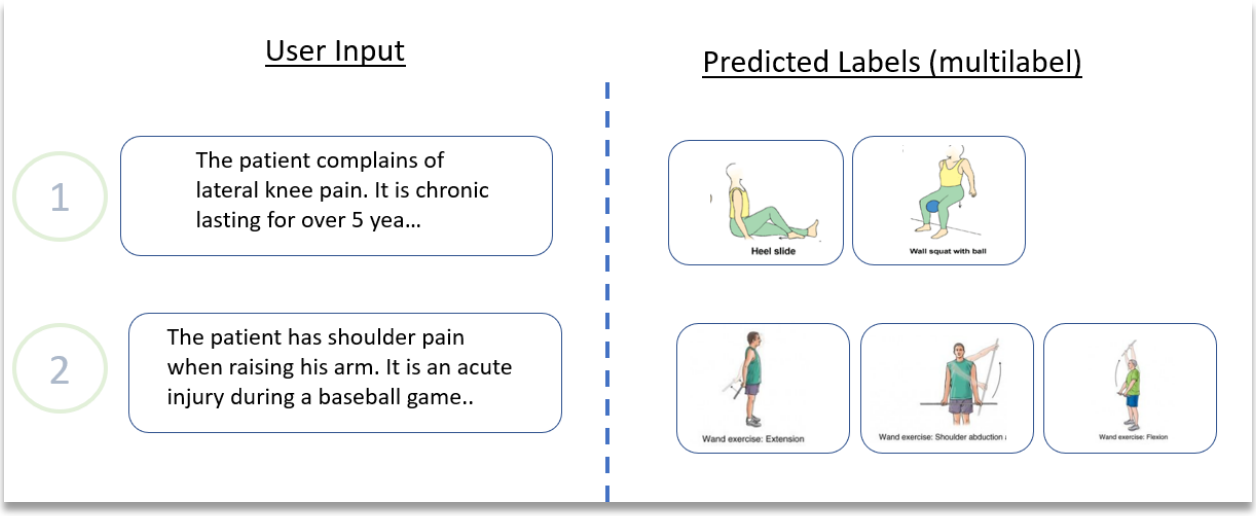
Also, is the exercise dataset comprehensive? How does this change over time? The reality is that not every possible exercise can be predicted, but over time as better exercises emerge, the classes will be updated to include it. As such, the application design allows for growth and feedback.

Choice of Data Labels

This data is already labeled, but as new data (exercises) are provided, it will be labeled appropriately according to the type of exercise that it is, by licensed physical therapists. The training data is already text data with defined labels (classes of exercises), so this is less applicable to this problem.

This is not a binary or multiclass problem, but a multi-label problem. It is more similar to the ImageNet competition in that a large number of classes exist, but different in that each prediction is a combination of classes. For example, one record may contain text data indicating a knee injury (like record #1 below), and the physical therapist provide 2 exercises, whereas a different physical therapist may have provided 5.

Below are some examples of the classes and labels. The labeling will be in accordance with common nomenclature mirroring the American Physical Therapy Association. Some examples include in example #1 below "Heel Slide" and "Wall Squat With Ball".



This labeling scheme allows for a large variety of exercises (classes) and thus can be differentiated for multiple recommendations (exercises for various ailments). It is easy to understand, and based on common APTA nomenclature, so it should line up with what physical therapists expect. So, rather than using a labeling scheme based on region of the body, or overarching exercise group (with less classes), this is more minute in detail and could give better predictions, a strength of the labeling scheme and thus application.

While this is a strength in potentially allowing more suitable exercises to a particular problem, it is a weakness in that it will potentially require a large amount of data to train the model successfully and likely create imbalance in the dataset (for edge cases). Acquiring more training data for modeling (beyond our initial estimate of 20K), may be costly through a third party, unless we solely rely on the data our application provides. A labeling schema as we have chosen requires ample data to be trained.

Due to the nature of this labeling scheme and application, it requires an expert to label the data, which is more expensive than if it were to be labeled on a platform, such as Figure Eight or AWS. However, it is not expected that new exercises are to be high in number.

Model

Model Building

The application will be built in-house using a cross-functional team with designated roles for the product owner, designer, software and data engineers, data scientists, and DevOps teams. The design will be iterative and will employ voice of customer surveys

and collaboration with stakeholders (physical therapists) to align product design with the market expectations and requirements, in an Agile methodology.

The model development will be in-house team, with the team of data scientists using opensource tools for model development. The model is likely to be a multilabel classification model, though multiple models will be developed and tested, with the model metric as F1. This model and data size is not overly complicated, so an inhouse team will work well to mitigate costs.

It is likely that model hosting/scaling will be through Amazon Web Services (AWS) due to security concerns, and the need to scale the product to potentially thousands of separate clinics as a full AI product offering (website/app). This platform also allows for A/B testing, monitoring, and all necessary tools for web application development and hosting.

Evaluating Results

The model metric will be the F1 score, since it takes into account recall and precision. This will assist with a more balanced model. Since this is a difficult model to benchmark, our first model will serve as the benchmark model and through development will rely on the train/validation/test F1 score compare and make improvements.

The data science team and AI product owner will pay close attention to the classes that are not correctly predicted, and that the dataset is balance (or artificially so). If one class does not train well, additional records will be acquired or created.

However, the model will also be monitored in deployment using the following metric:

- percentage of predicted exercises of the total exercises provided to the patient

For example, if the AI product recommends 10 exercises, and the therapist provides 3 of these along with 2 not predicted (for a total of 5 exercises given to the patient), then the metric is 60% (3/5). This will indicate how many are selected from the predictions, where the higher the better.

Even though our model may produce strong accuracy on F1, it needs to align with the business need. If the products are not being selected, then there is a disconnect that needs to be addressed. Both metrics (F1 and the custom metric) allow for finding lapses in the model or correlation with business outcomes.

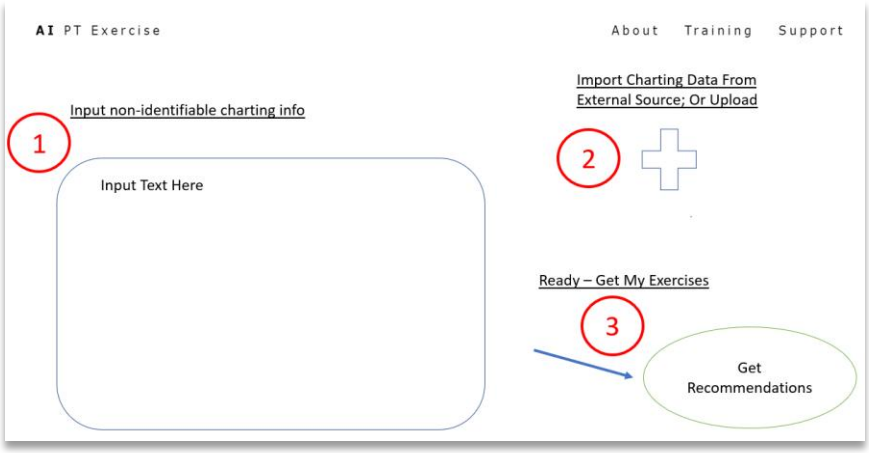
Minimum Viable Product (MVP)

Design

The product will be a standalone subscription-based web app. The minimum viable product as follows, and is hosted on a scalable, secure platform such as AWS.

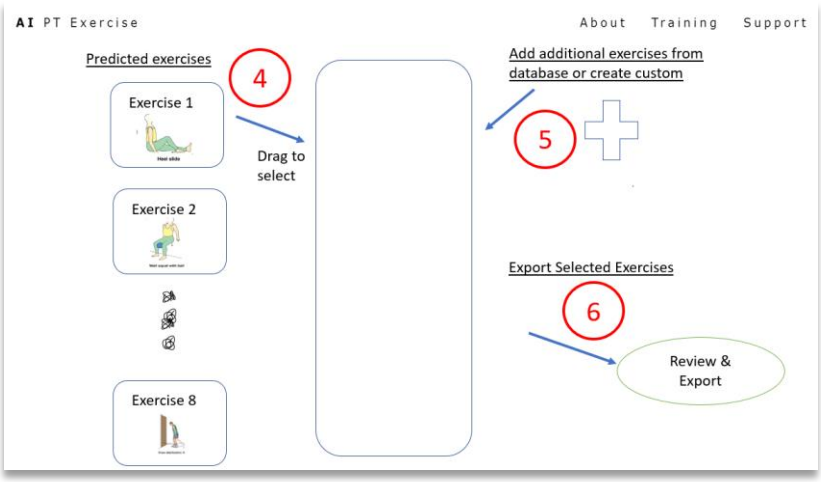
There is a simple user interface to (1) input text data or (2) import data/upload from an external source. Once completed, the user (3) selects "ready" and the application begins to predict and move to the next site.

MVP: User Input Interface



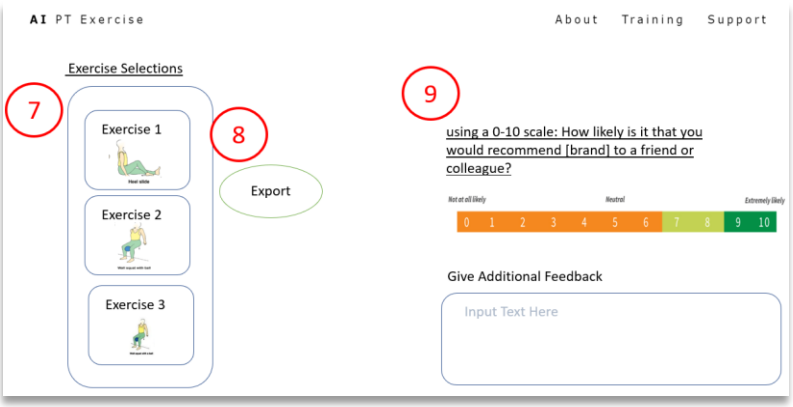
The user is directed to this interface. As shown below, the predicted exercises are provided, with the option to (4) drag to select the predicted exercises and (5) import/create new ones. Once completed, the user (6) selects review/export.

MVP: AI Product Selection Interface



MVP: AI Output & NPS Feedback Interface

In (7) the user reviews and export a data product of their choosing (a PDF, email, printout,…). In (8) we request user feedback in a simple question (using NPS) with an option text-box to elaborate. This gathers information related to our success metric to customer experience.



On a side note, the website contains options for feedback, rating of each exercises, training materials, etc. However, since this is an MVP, this is the minimum it would do to be viable and tested in the market.

Use Cases

This is specifically designed for a physical therapist or assistant. While working with a patient and gathering information, the notes are input or imported into the application, which drive exercise recommendations and allows the physical therapist to quickly review/select exercises or add manually. This should allow more time with the patient,

and less “searching”. The final product can be a printout to review with the patient, and/or an email with a pdf. A few use cases:

- “As a physical therapist, I want the app to recommend exercises for patients so I save time at the computer and can spend more time with the patient.”
- “As a physical therapist, I want the app to create easy to use products of exercises for patients to use.”
- “As a physical therapist, I want the app to recommend exercises for patients so I can consider other exercises to provide, or learn new exercises myself.”
- “As a physical therapist assistant, I want the app to recommend exercises to assist me in preparing exercise routines for the physical therapist (who reviews).”

Roll-out

The rollout will initially be limited to a few clinics to test the product and gain valuable feedback. It is our intent to test the MVP as soon as possible, to ensure it is viable in the market. Obviously, user experience (physical therapist) is incredibly important, so will be collaborative with design.

Upon user acceptance, and good predictive accuracy based on our business metric, iterative improvements will be made. Final rollout to a larger market will be completed once a more polished product (and supporting material and training information) is ready with market positioning.

Post-launch will include monitoring of the AI product and requesting feedback from random clinics (what they like or don’t like, potential problems or opportunities). If there is feedback received, it will be reviewed by the product owner.

Post-MVP-Deployment

Designing for Longevity

Long-term improvement relies on a growing dataset, and this has been built into the design of the application. In the design sketch, there is an option (5) to “add additional exercises” or “create”. This will not only allow for the exercise data to potentially grow but will also align the specific user input with selected exercises. As such, our training data increases with time which will allow for improvements over time.

Real-world data with vary considerably across the United States, so it is important to allow our product to learn from new data. A/B testing will be employed with iterative model development to test if a new model should replace the current model (using 20%

of data on the "challenger" model. As the model is changed, versioning will be documented via GitHub. Test cases will be provided to the model from the original dataset, to measure performance over time based on the original dataset and also the new dataset, as models are trained with new data.

If the event occurs where this product appears to be highly biased to a certain specialty of physical therapists and poorly for others, this will be investigated and a possible new application to create and offer alongside this product.

Monitor Bias

Bias will be combated with the growing database and with models routinely, iteratively designed. The in-house data science team will review (1) predictions that are not selected and (2) "manual entries" as time passes. In a way, this review of the False Positives and False Negatives will allow us to see where the model is failing, or perhaps needs to be balanced. Monitoring on AWS is fairly routine, and if the business metric begins to fall or increase beyond a reasonable percentage (5%), then it will notify the team to review.

Guarding against data bias due to imbalanced data input. For instance, a knee injury is more common than a toe injury, so imbalance may be in the training data as a result. These edge cases need to be capture for a robust product. This will be remediated if it affects the business outcome via rebalancing the dataset, or including more examples in the training data. This will also involve training a new "challenger" model, A/B testing, and possible deployment. If AI development for these edge cases is difficult, alternative tools will be created for these edge cases.

Meetings with stakeholders and existing users should provide overall feedback to the product, and allow the team to improve the design, add needed features, or expand the product offering.

Resources

- https://blog.marketresearch.com/u.s.-physical-therapy-clinics-constitute-a-growing-34-billion-industry
- https://www.microsoft.com/en-us/trust-center/cloudservices/health
- https://scikit-learn.org/stable/modules/multiclass.html
- https://www.zdnet.com/article/launching-oracles-covid-19-therapeutic-app-the-back-story/
- https://peoplepulse.com/resources/useful-articles/net-promoter-score-nps-implement/
- https://www.apta.org/