System Design and Ethical Considerations for Project Proposals

Project Title: DL Alpha – Student Journey Insights
Team Members: Jason Wheeler and Sergio Rio

A. Machine Learning System Design

From the System design perspective let's consider the three aspects:

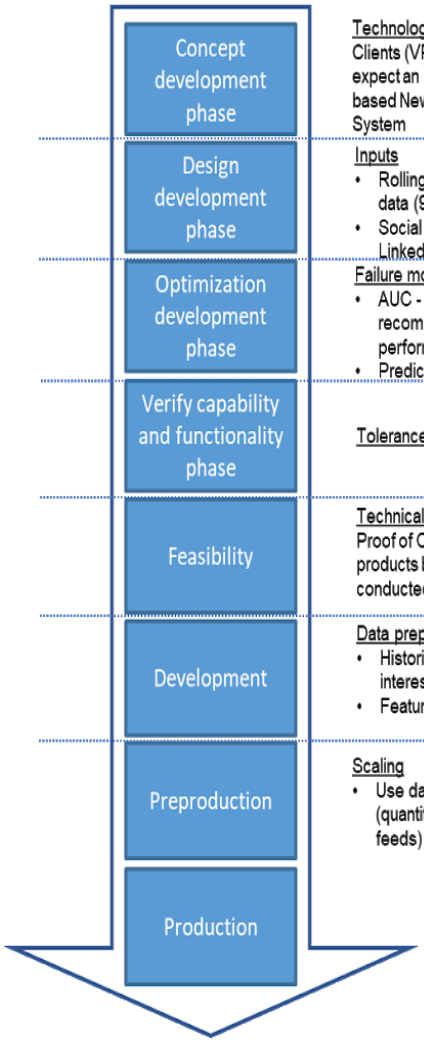
1. Data:
   Database includes approximately 300k individuals. Each individual's record has activity like opening marketing emails, viewing web pages, and registering for events. Some also have data on their location and job title, and newsletter subscriptions: The Batch (a weekly content newsletter) Events (events put on by DL and members of their community) and Course Announcements (used when DL releases a new course).
   The Contact database is stored in HubSpot. The CRM includes individual IDs with marketing and engagement information including fields like "Last Page VIewed", "Last Marketing Email Opened" and "Last Event Attended".
   When a user that is known to DL.ai visits the site, Hubspot tracks this with cookies but does not keep a record of every page visited. Instead, it updates the field "Last Page Viewed" at the end of a web session.
   All fields named "Last …." that update based on user behavior (like registering for an event, viewing a marketing email or a webpage) have a field history csv, separate from the main contacts CSV.
   When registering for a webinar, speaker, or other event, DL.ai asks about their course history with DeepLearning. This is the source of identifying which students have paid for a course. The field is called "Which of the following online courses have you taken from deeplearning.ai?" This field also has the history of updates in a separate CSV.
   Individual Data Sets:
   - Entire contact DB from Hubspot with location, jot tible, etc for those we know. Includes engagement data including:
   - last marketing email
   - last page viewed
   - last event attended
   - with corresponding date/time
   - Field history for "last marketing email name", open date
   - Field History for "last page viewed"
   - Field History for "Last event attended"
   - Field History for "Which of the following courses …"

   Journey level features will need to be extracted for each student followed by pre-processing the data to categorize students based on the numbers of paid courses taken. The goal is to identify key analytics, such as: majority of students take what number of paid courses?, what are common behaviors that inhibit taking more courses? what can be recommended pathways to enable students to take more courses?
2.

Insights - System Life Cycle End-2-End Product Design

Design for Six Sigma: Roadmap for Successful Corporate Goals, Salman Taghizadegan, in Essentials of Lean Six Sigma, 2006

3. Process (Models, iterations)



Technology Concepts
Clients (VPs of Sales and Operations) expect an innovative Machine Learning-based New Product Recommendation System

Goals
Maximize the return over the investment and revenue growth from new products

Constraints
- New system must be released to production in 3 weeks
- 3 years of past customer journey history
- 1 Data Scientist (part-time) and 1 ML Engineer (full time)
- Dedicated computing environment with GPUs/TPUs

Inputs
- Rolling window of historic customer journey transactional data (9 months for training and 3 months for validation)
- Social media feeds from Facebook, Instagram and LinkedIn

Outputs
- Recommendation of products to be launched (0: not recommended 1: recommended

Flow Diagram(*)

Data Privacy and Bias
Customer and PII will be removed before used by model data pipeline

Failure mode effects analysis (FMEA)
- AUC - ROC curve to measure recommendation (classification performance)
- Prediction accuracy/recall/f1-score

Random baseline: run random classifications on historical data
Human baseline: gather performance of existing R&D and Marketing teams on new products recommendations
Simple heuristic: gather voice of the customer and customer surveys

Tolerance design analysis

Technical Feasibility
Proof of Concept with top 100 products by sales volume will be conducted

Economic and Operation Feasibility
confirmed by project budget and available company resources and specialized personnel

Data preprocessing & representation
- Historical data will be pre-processed to extract features of interest
- Features will be analyzed using Random Forest and OLS

Model Selection
1. Iterative process to select best model:
2. TPOT(Auto ML) for classifier based on quantitative historical data
3. Co-training for classifier based on social media feeds

Training & Hyperparameter Tuning

Feature Engineering

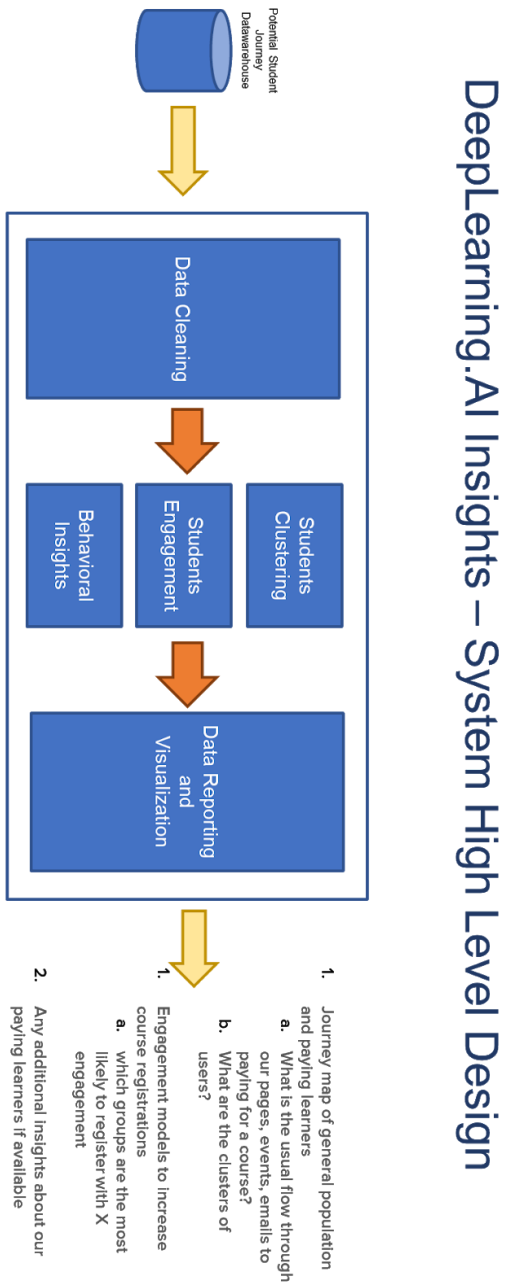
A/B Testing

Scaling
- Use data parallelism for training model pipeline (quantitative data) and co-training (social media feeds)

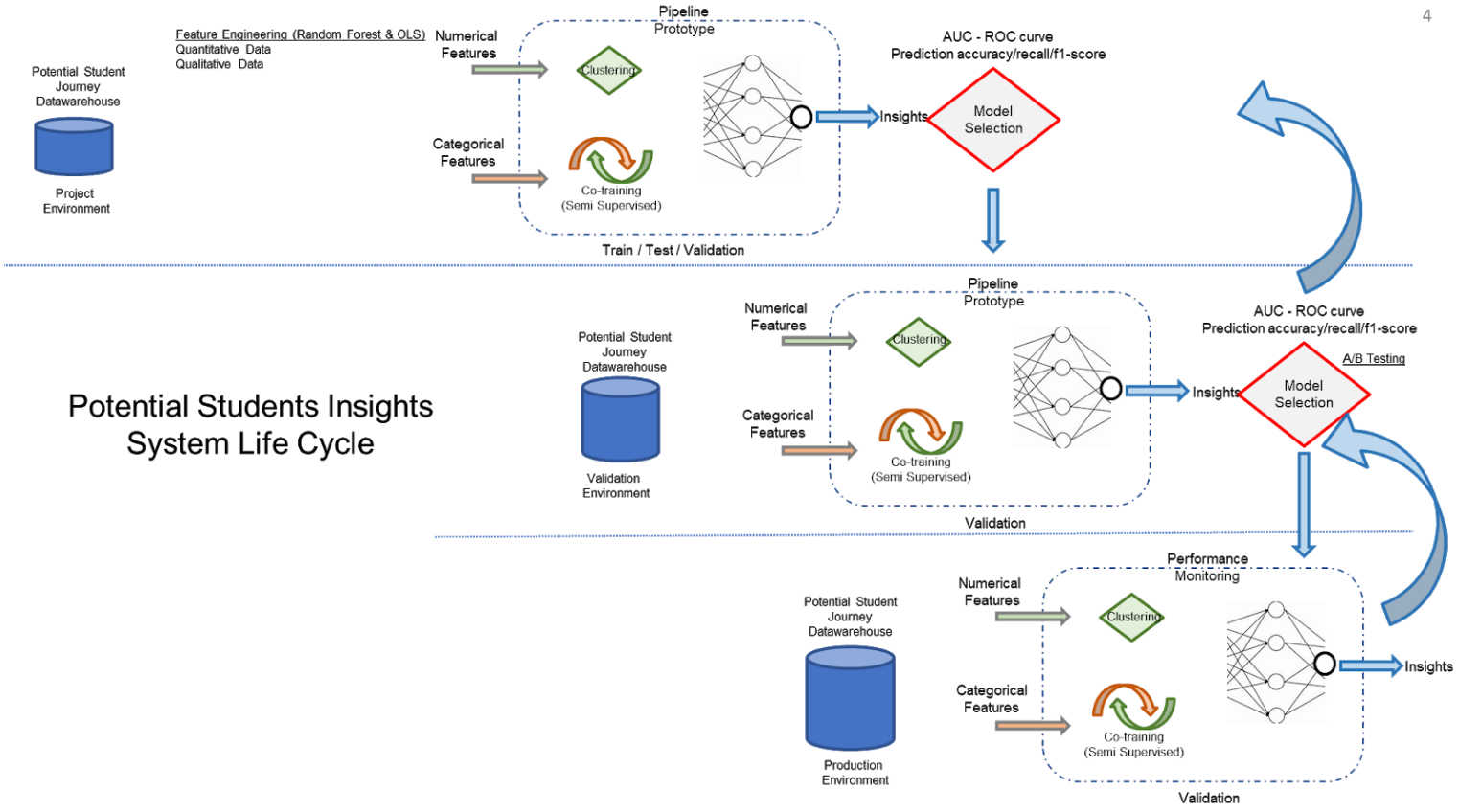
Serving
- Run experiments with Sales and Operations planners to confirm end user experience, predictor behavior, boundaries, errors, results interpretability and model explainability
- Conduct ablation studies -- removing each component while keeping the rest -- to determine the efficiency of each component. Find components whose removals eventually don't significantly reduce the model's performance but significantly reduce its complexity.
- Check model for propagating any gender and racial biases (if only if gender and race are features used)

4. Outcome (output and recommendations)

   What are the system design considerations for your deployable ML model. Describe the iterations, delivery formats and limitations you may face and some solutions to overcome the limitations



Potential Students Insights System Life Cycle



B. Ethical Considerations

Are there any ethical considerations of your project? Consider the data source, the intended outcome, and/or the eventual use cases.

- Datasets are property of DeepLearning.ai and will be kept confidential and used solely for the purpose of this project. Upon project completion, the project team will delete copies of the data sets from computers and cloud storage.
- All PII (Personal Identification Information) needs to be removed or masked (e.g., names, addresses, e-mail addresses)
- Features that could potentially drive bias, such as gender, race, location and nationality will be identified and analyzed early in the project. Potential model bias (e.g., gender, nationality, etc.) needs to be addressed by design and discussed with project sponsor in case there's any impact on the final product or deliverables.
- Models and insights need to be tested against selected features
- Analysis, insights, reports and visualization you will be neutral by design
- Our models will be static and could potentially produce biased insights/findings as a side effect. Fail safe checks must be implemented "a priori" as much as possible.
- When demonstrating the model, we must be vigilant

- Did you modify anything about your plan based on these considerations?
  We will add specific tasks/activities to our project plan

- Can you anticipate any issues that might arise during the process?
  Yes, at this moment we anticipate we'll probably need additional customer information from Deeplearning.ai that can't be disclosed (PII).