**Name:**  DSBA Student

**Date:** 3 April, 2022

**Project Title:** Proof of Concept ML Pipeline in Google Vertex AI

**Mentor and Company:** Mentor Name, Company Name

**Period Covered**: 21 March – 1 April 2022

**Project Objective:** Build, test, analyze viability, and improve performance of a proof-of-concept start-to-finish machine learning pipeline.

**Progress Summary:**

Work Completed.

* **Task 3.2.** Make changes to `tx-prod` service to call the cloud function (**COMPLETE / ADDITIONAL FEATURES ADDED)** 
  + **Summary:**
    - Completion reported last sprint.
    - Further testing at the tail end of last sprint and beginning of this sprint revealed some additional modifications and refinements needed, which has further delayed subsequent tasks.
    - The cloud function now runs 50% faster than it did two weeks ago
    - The `tx-prod` service can now utilize the cloud function and Vertex endpoint during local development using Tilt
  + **Lessons Learned:**
    - Using caching (e.g. the TTL\_cache class from the cachetools library) on any functions that retrieve auxiliary data or authorization tokens can significantly decrease processing time of cloud functions.
    - As a best practice, instantiate classes prior to the GCF’s entry point. This way, one class instance can remain live for the entire life of the GCF, obviating, for example, the creation of new authorization tokens with each new instance.
    - The urllib3 python library’s Retry() class does not include the ‘POST’ method by default; it must be added explicitly in the allowed\_methods argument.
  + **Risk & Mitigation:**
    - Risk: Data lost due to overwhelming the GCF without a sufficiently resilient retry/resubmission structure
      * Mitigation: To prevent loss of incoming messages—mostly due to 429 HTTP errors—we increased the maximum allowed number of GCFs, increased the CPU count on the Vertex virtual machines (VMs), and modified the retry strategy.
* **Task 3.3.** DMI-184: Assess ground truth data (**IN PROGRESS / DELAYED)** 
  + **Summary:**
    - The Analytics off-site from 24-25 March delayed progress on this task
    - The data has been retrieved and the key summary statistics computed
    - A full manual re-labeling of ground-truth data will be needed
      * The data now contains 80% more vendors than previously, for whom the requisite category definitions are not readily available
      * The previous labeling process can be modified with assumptions, and used as a baseline to show how the data and the model’s performance may have drifted over the past year
  + **Lessons Learned:** 
    - The sales\_type field has changed since the last time the model was trained, making the previous labeling process no longer tenable without some assumptions and modification.

Work Scheduled Next Sprint.

\*\* The quarterly product increment (PI) planning conference runs from 5-7 April, which will limit progress on scheduled work for this sprint.

* **Task 3.3.** DMI-184/CP-1011: Assess ground truth data
  + Develop rules-based model for ground truth labeling through manual inspection, research, and heuristic development.
  + Recompute summary statistics based on new ground truth labels
* **Task 3.4.** DMI-183: Define, compile, and deploy the `is\_x` pipeline in Vertex AI
  + Build components to:
    - retrieve data,
    - engineer features,
    - train the model,
    - evaluate performance, and
    - deploy, conditional on meeting the performance threshold
  + Store reusable components in co-vertex-utils
  + Write and pass unit tests for all reusable components
* **Risks & Mitigations**
  + Risk: Rules-based labeling can become complex and time-consuming as the definition of the true labels is necessarily a best approximation
    - Mitigation: For now, allow the assumption that matches to the Industry Product Master are definitively X product; relegate any further refinement in this respect to another ticket / future work.
  + Risk: Uncertainty due to troubleshooting on the pipeline construction and compilation, acknowledging Google’s documentation is sometimes poor.
    - Mitigation:
      * Coordinate closely with coworkers working concurrently on other pipelines to maximize sharing of work and reusing of component structures.
      * Limit work on this ticket to building the pipeline that previously trained and can retrain the `is\_x` model; relegate substantive *improvements* to a subsequent work package.

**Work Package Updates:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **WP** | **Work Package Title** | **Scheduled Dates** | **Status** |
| **✓** | **~~1.~~** | **~~Onboarding~~** | **~~10-Jan – 11-Feb~~** | **Complete** |
| **✓** | **~~2.~~** | **~~CP-856: Test performance of `is\_x` model on Vertex Endpoint~~** | **~~17-Jan – 18-Feb~~** | **Complete** |
|  | **3.** | **CP-857: Build a proof-of-concept ML pipeline in Vertex that can effectively replace the current is\_x data pipeline** | **21-Feb – 15-Apr** | **In progress** |
| **✓** | ~~3.1.~~ | ~~CP-988: Create a Google Cloud Function to effectively replace the `ml-x` service~~ | ~~21-Feb – 11-Mar~~ | Complete |
| **✓** | ~~3.2.~~ | ~~CP-989: Make changes to tx-prod to call CF~~ | ~~14-Mar – 18-Mar~~ | ~~Complete~~ |
|  | 3.3. | DMI-184/CP-1011: Assess `is\_x` ground truth data | 21-Mar – 8-Apr | In progress |
|  | 3.4. | CP-857/1020: Build POC Vertex pipeline for the `is\_x` model | 11-Apr – 15-Apr | Backlog |
|  | **4.** | **Explore options for improving `is\_x` prediction performance. ~~and expanding to additional product categories.~~** | **18-Apr – 29-Apr** | **Backlog** |
|  | 4.1. | DMI-123: Spike: Evaluate `is\_x` performance and recommend improvements | 18-Apr – 22-Apr | Backlog |
|  | 4.2. | DMI-382: Update `is\_x` pipeline based on findings from DMI-123, and retrain model | 25-Apr – 29-Apr | Backlog |
|  | ~~4.2.~~ | ~~Explore model expansion~~ | ~~21-Apr – 29-Apr~~ | ~~Backlog~~ |