Team ODB

Sprint 3 Report

Prepared by:
Michael Dykema
Tom Lobbestael
Emily Linderman
Jacob Johnson

Intended Progress

Michael: The first task I intended to complete was to find a suitable dataset to use temporarily for building and training our recommendation model. Ideally this would be a Google Analytics dataset as this would more closely resemble the format of the actual data, but any dataset good for making recommendations would work for us at this point. The second task I planned on completing was to attach the API to the recommendation model. This would include the research on how to connect to the SageMaker model and return the results.

Tom: For this sprint I intended completion of 1.0 of the SageMaker recommendation engine. That is to say using a sample dataset found online to make recommendations. Following this if extra time was available the tentative plan was to either work on hooking the engine to the API and data lake or to assist in the data lake as needed.

Emily: This sprint I intended to create the CloudWatch Dashboard through the AWS Console to monitor our API.

Jacob: This sprint I intended to research Kinesis Firehose with the purpose of adding streaming of new data into the data lake skeleton, as well as work on the transformation stage of the data lake that would aid the machine learning algorithm.

Progress Reflection

Michael: I was able to find a few potential datasets that matched our use case somewhat, but none were a great fit. Turns out people don't like to post detailed user data (even anonymized) for free to the internet. With that completed I was able to figure out that when building and deploying a SageMaker model, there is an endpoint created that will return results (recommendations in our case) and I was able to build a Lambda function to retrieve those results. The Lambda/API will need to be updated once we have the finalized formatting for content cards from ODB, but the initial task of connecting to the recommendation engine is complete.

Tom: Being honest I didn't accomplish near as much as I should have this sprint, mainly due to procrastination during break. I did manage to acquire the dataset that Micheal found and make some minor changes to the SageMaker code.

Emily: Due to this sprint containing a school break, not as much progress was made as was initially planned. I successfully figured out how to make the dashboard and looked at some of the widgets available to add to the dashboard, but was unable to implement the official dashboard due to lacking permissions.

Jacob: I spent much of the first half of the sprint reading up on how Kinesis Firehose works and how it will go hand in hand with LakeFormation. After that I was able to add an implementation of new data streaming to the data lake skeleton which will be ready to have the Amazon resource names copy pasted into it when they are created. I did not manage to work on the ETL stage in the data lake past just streaming data in, because the research and implementation of Firehose took so much time and because the format of the ODB data is yet unknown.

Problems Encountered

Michael: Aside from the struggles of finding a good dataset for user analytics, I didn't run into any major issues. The only thing needed to finish the API will be the formatting for the content cards which I need to get from ODB and details on any specific error handling they desire.

Tom: The biggest issue that I faced this sprint was actually acquiring the dataset that we decided to use. It was hosted through Google BigQuery and thus not really meant to be exported. To achieve the export it took a good amount of research and time. Other than that break led me to procrastination which led to issues in progressing on SageMaker.

Emily: The only issue I encountered over the course of this sprint was the lack of permissions in the AWS Console, which meant I could not implement the dashboard, as the dashboard can only be created through the Console.

Jacob: I ran into an issue with conflicting commits because I forgot to pull down new changes from Michael before doing my own local work but other than that I had no issues.

Projected Progress

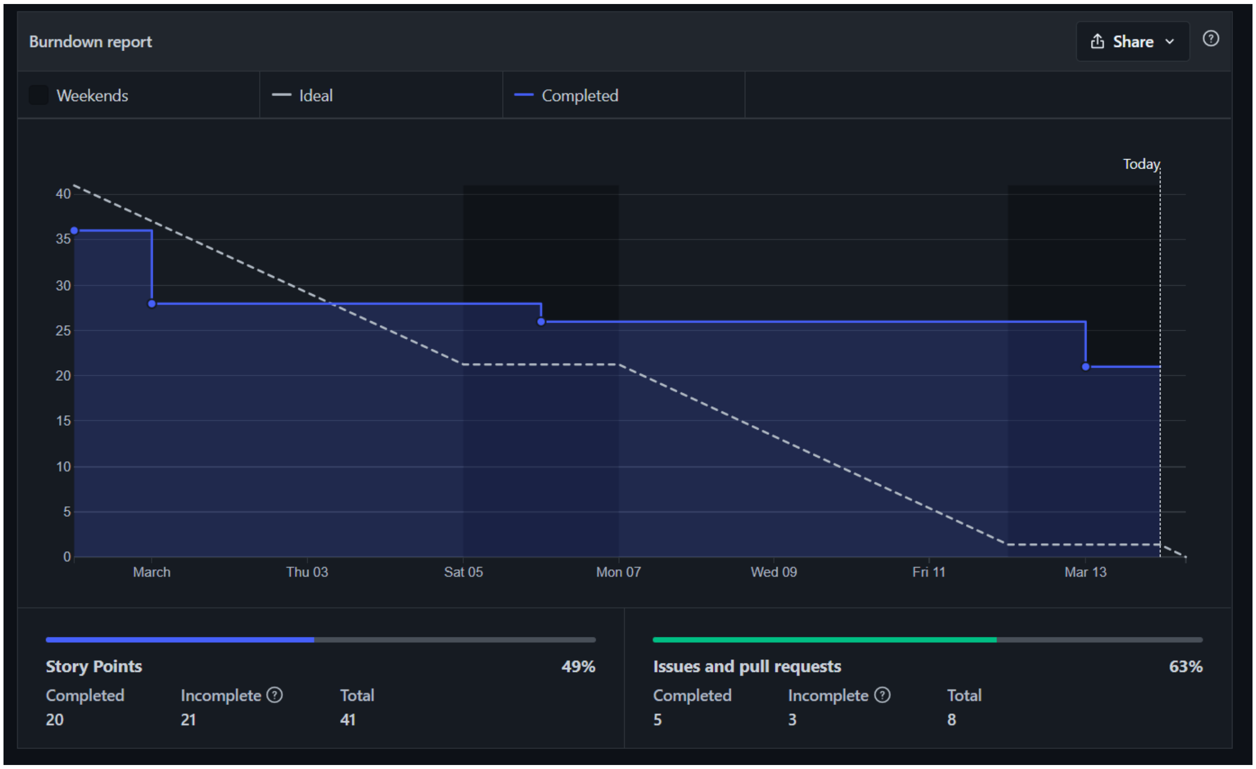
Michael: Next I will be finalizing the API for returning complete content cards. Additionally, I will be working on the recommendation model. I will probably need to do a little bit of research into TensorFlow before I can get started, but I anticipate being able to help with the training and potentially even working on a second model to compare and analyze. There is the possibility that we will need to do some categorizing of ODB web pages to return results and so I will work on that if necessary.

Tom: Next I plan to complete v1.0 of the recommendation engine (over the next few days) and after that I plan to work on hooking it up to the API and data lake. After those are completed, if ODB has provided us with their data and the lake is populated then I will work on honing the engine and if not I will work on DataDog.

Emily: Next sprint I will be properly implementing the dashboard, as we hope to have the appropriate permissions as soon as possible. Once that is finished I will work on the other CloudWatch aspects we will need for the project.

Jacob: Next sprint I will work on creating the actual resources in AWS, such as buckets, and plugging their metadata info, such as ARNs, into the data lake skeleton.

Burndown Chart



Story Points 49%

Completed	Incomplete	Total
20	21	41

Issues and pull requests 63%

Completed	Incomplete	Total
5	3	8

As can be seen by our burndown chart we are still working on having a more accurate estimation of what we can actually achieve during a sprint. Along with this one of our incomplete issues is nearly completed and will be done within a few days. Overall, this sprint we didn’t achieve all that we wanted but did get a lot done. For future sprints we should have a generally good idea of what amount of work we will actually be able to complete and future burndown charts should reflect that.

Teamwork Reflections

Overall, the level of teamwork and communication was improved over this last sprint compared to the first two sprints. Despite the fact that this sprint occurred over GVSU’s spring break and our project sponsor was away for most of the sprint, the team still met and went over progress, problems, and goals.

While members were working on tasks individually, there is probably room for improvement in terms of when an issue is encountered asking for help from team members or pivoting off the current task to work on something that does not have any blockers.

Conclusion

It was at this point in the project that we originally projected to have a working ML recommendation engine, however we are not quite there yet. The goal is to be at that point by the end of this next sprint. The data lake and API are in good places where they are waiting on specifications and work to be done on ODB’s end. CloudWatch monitoring is making good progress but the dashboard for the visualization of this monitoring is awaiting the necessary permissions to create it. The recommendation engine itself needs a bit more work but with at least two team members working on it this next sprint the first model should be finished shortly.