Assurant Georgia Tech Practicum Summer 2024

Muhamad Imannulhakim, Aditya Hedge, Joseph Lee, Sarthak Khare, Kyrylo Kobzyev, Priyanka Singh, Yang Lu

ASSURANT®

Georgia Tech

Table of Contents

Georgia Tech

Introduction to the Problem - Part 1

Microsoft Azure

- Assurant, Inc. is a global provider of risk management products and services with headquarters in Atlanta.

- Assurant uses Microsoft Azure's Cloud Services for its operations, and since Microsoft manages users' sensitive information(on Assurant's behalf), they must provide structured documentation detailing what they're doing to protect that information.

- These are done through SOC® examinations. SOC stands for System and Organization Controls. It's a type of examination geared toward entities that provide services directly related to a user's control systems, like SaaS companies, financial reporting organizations, data centers, and payment processors.

Georgia Tech

Introduction to the Problem - Part 2

- SOC 1 focuses on a service organization's controls related to financial reporting. Entities utilizing these service organizations might request a SOC 1 report to assess how the controls impact their own financial statements. This is crucial for both the entities and the CPAs auditing their financial statements.
- SOC 2 evaluates a service organization's controls according to five criteria: security, availability, processing integrity, confidentiality, and privacy. A wide range of users might request this report to obtain detailed information and assurance about the service organization’s controls concerning 1) the security, availability, and processing integrity of the systems used to process users’ data, and 2) the confidentiality and privacy of the information processed by these systems.

Introduction to the Problem - Part 3

- Since Microsoft has measures in place to ensure SOC compliance, certain actions attempted by Assurant's IT team to Create, Enhance or Modify Existing Data Pipelines and Assets are rejected on account of being non-compliant.

- Assurant as Organization wants to prevent any requests or actions performed that will be rejected by Microsoft's Data Policies.

Georgia Tech

Potential Solution

- Our solution to this problem is to develop a model that can classify proposed actions as potential failures. This will reduce the time needed to identify, diagnose, and repair incidents in IT systems using machine learning.
- We are focusing on:
  - Developing predictive analytics
  - Conducting causal analysis
  - Implementing automated remediation
  - Establishing a CI/CD pipeline with automated logging and a learning loop
- The expected deliverables include:
  - ML models for predictive analytics and event correlation
  - Automated remediation systems
  - An integrated CI/CD pipeline
  - An automated logging framework
  - Comprehensive documentation and training materials
  - Performance reports
  - Future improvement plans

Georgia Tech

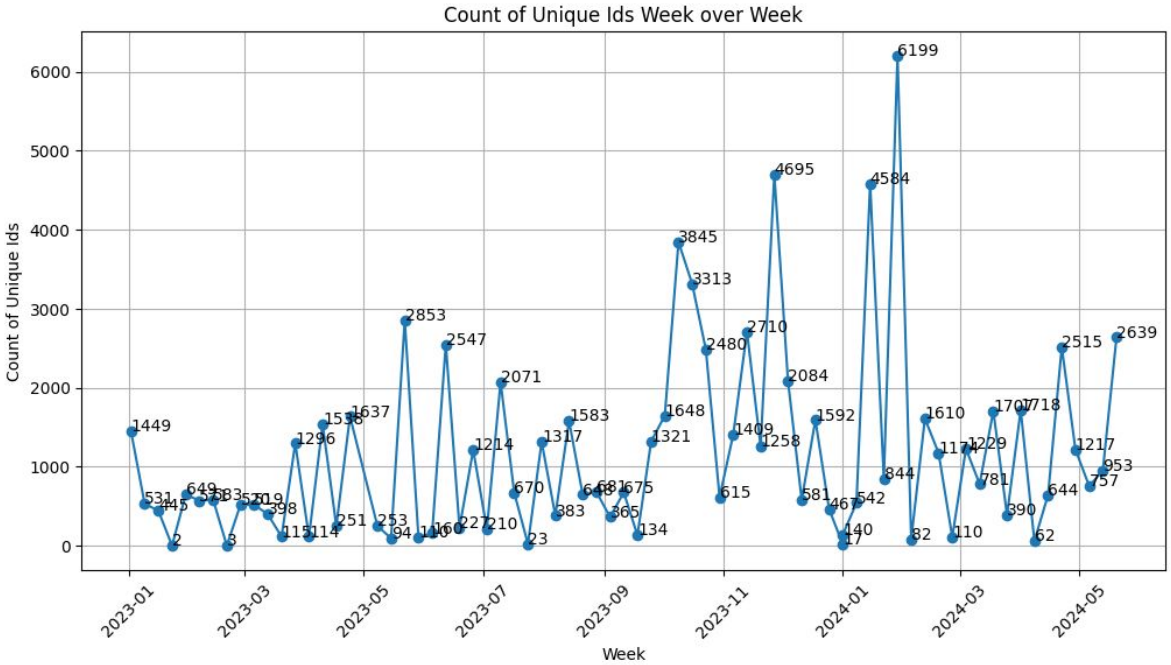
Data

- The Data we have been provided are log entries provided by Microsoft to Assurant's IT Team.
- The data itself consists of ~250K entries for a period spanning 18 months.
- Since the data is in a Standard form that Microsoft provides to all its clients requesting their Logs, the data itself is fairly clean and standardised since Microsoft is one of the preeminent providers of Cloud Services to organizations the world over.
- Additionally documentation regarding the data is available easily available online on the Microsoft Azure Help Section.

Exploratory Data Analysis

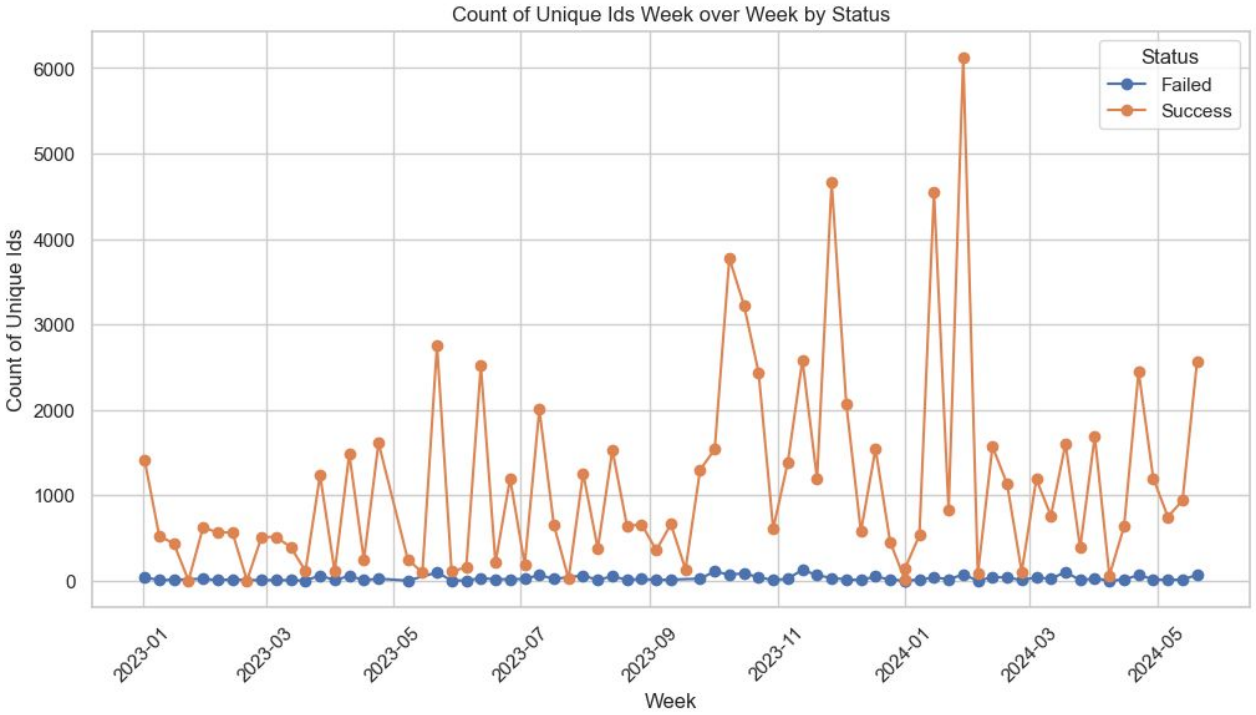
- We started with an EDA to understand the data, here are some major findings:
  - About 20% of the data provided was duplicated, and the duplicate rows were removed from the data before further analysis.
  - More than 50% of the data is from May of 2024, alone. Specifically, between the 26th and 31st of May. The data will be divided into - before 25th May and after 25th May for further analysis.
  - ~4.8% of “Execute” log entries are failures in the prior timeframe.
  - All Failures are tagged as either “Pipelines” or “Release”.

Exploratory Data Analysis - Chart 1



- The count of Log entries week over week shows spikes October 2024 to January 2024

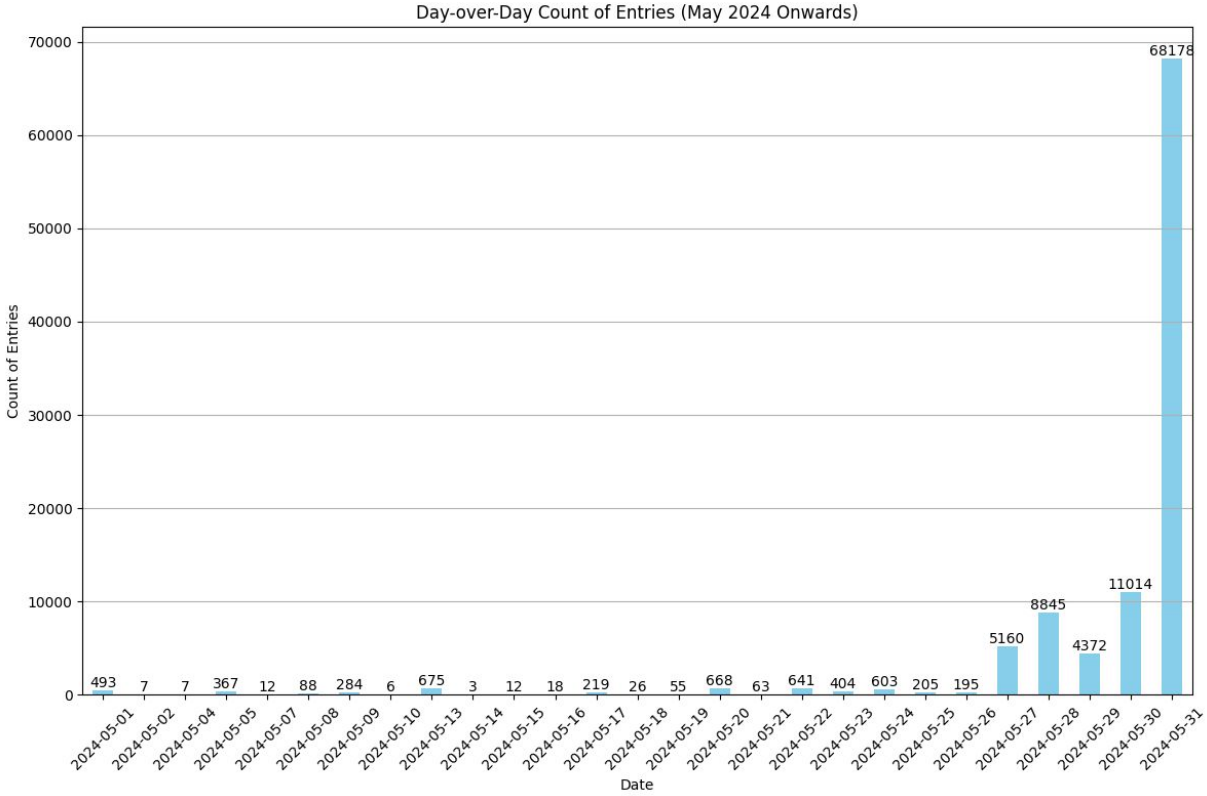
Exploratory Data Analysis - Chart 2



- However there is no spike in Failures during the same time, and they remain relatively stable over time.

Georgia Tech

Exploratory Data Analysis - Chart 3



- We need data from June to understand if there have been any other stark variations in the log entries recently, or if any underlying issue has been solved.

Georgia Tech

API Development

- API Functionality
  - We deliver the results of our prediction model through an API.
  - Provide reasons for task failures.

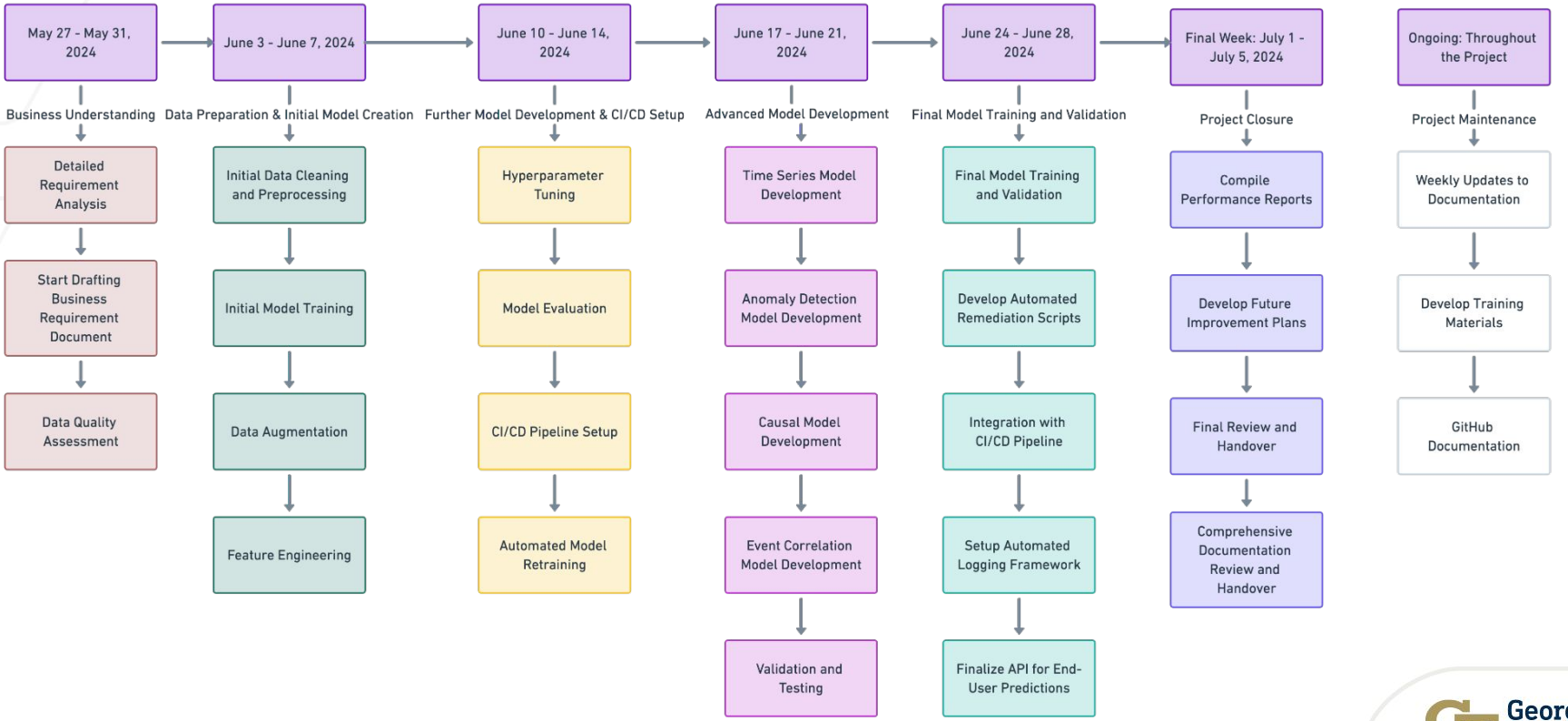
- Current Status
  - Initial API has been developed.
  - Tested response both locally and publicly (using Ngrok for tunneling).

- Future Goals
  - Aim to have a real-time API service and potentially implement containerization.
  - Obtain results and provide reasons when new log data appears, guiding developers for fixes.
  - Utilize FastAPI's built-in Swagger for API documentation.

Current Status

- We are completed an Exploratory Data Analysis
- We have set up a skeleton framework for the API
- We are working on Feature Engineering and building first drafts of the models/classifiers

Task Flow Diagram



Georgia Tech

Detailed Timeline and Task Distribution - Part 1

Week	Phase	Tasks	Assignees
May 27 - May 31, 2024	Business Understanding	Detailed Requirement Analysis	All
		Start Drafting Business Requirement Document (BRD)	All
		Data Quality Assessment	All
		Exploratory Data Analysis (EDA)	All
June 3 - June 7, 2024	Data Understanding and Preparation	Initial Data Cleaning and Preprocessing	Aditya, Iman, Kyrylo
		Advanced Data Cleaning and Preprocessing	Aditya, Iman, Kyrylo
		Feature Engineering	Sarthak, Priyanka, Lu
		Data Visualization	Sarthak, Lu, Joseph
		Finish Drafting Business Requirement Document (BRD)	All

Georgia Tech

Detailed Timeline and Task Distribution - Part 2

Week	Phase	Tasks	Assignees
June 10 - June 28, 2024	Model 1 Development - Predictive Analytics	Model Selection and Initial Training	Aditya, Iman, Joseph
		Model Evaluation and Tuning	Aditya, Iman, Joseph
		Initial API Development for Model	Aditya, Iman, Joseph
	Model 2 Development - Predictive Analytics	Model Selection and Initial Training	Priyanka, Sarthak, Lu, Kyrylo
		Model Evaluation and Tuning	Priyanka, Sarthak, Lu, Kyrylo
		Initial API Development for Model	Priyanka, Sarthak, Lu, Kyrylo
	Model 1 Development - Causal Analysis & Event Correlation	Causal Model Development	Aditya, Iman, Joseph
		Event Correlation Model Development	Aditya, Iman, Joseph
		Validation and Testing	Aditya, Iman, Joseph

Georgia Tech

Detailed Timeline and Task Distribution - Part 3

Week	Phase	Tasks	Assignees
July 1 - July 5, 2024	Automated Remediation & CI/CD Pipeline	Develop Automated Remediation Scripts	Iman, Lu, Joseph
		Integration with CI/CD Pipeline	Joseph, Priyanka, Aditya
		Setup Automated Logging Framework	Joseph, Kyrylo
		Finalize API for End-User Predictions	Sarthak, Lu
Ongoing: Throughout the Project	Documentation and Training Materials	Weekly Updates to Documentation	All
		Develop Training Materials	All
		GitHub Documentation	All
Final Week: July 1 - July 5, 2024	Performance Reporting and Future Planning	Compile Performance Reports	All
		Develop Future Improvement Plans	All

Georgia Tech

Thank You!

Georgia Tech