

PwC Internship Project and Tax Technology Insights

CS4991 Capstone Report, 2024

Irvine Madenga
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
xth3kz@virginia.edu

ABSTRACT

The main objective of my internship experience was to find a more efficient way to parse through tax documents and extract data from them. The software my team was using to parse data was a third-party service that posed financial and internal issues. My team was able to leverage Azure Document Intelligence Services to more efficiently extract data from forms. Through this experience I was able to gain valuable insight into the agile methodologies used to solve problems in corporate tech environments and seamlessly work in a team on a large-scale project. At the end of the summer, we were able to implement Azure Document Intelligence and extract fields from tax forms into our databases. Some insights for the future regarding this project include an expansion of the service, full streamlining of the data extraction and potential AI chatbots that would be able to help clients analyze the tax forms they submit.

1. INTRODUCTION

During the summer months of 2024, I had the pleasure of working as a Software Engineering Intern in Washington DC. My work was centered around a specific tax tech team in the company that was developing and maintaining a web application that allowed clients and partners to be able to access and process tax documents that the company was providing service for. Through

this work I have been able to explore the leveraging of AI tools in professional settings. This was particularly important since I got to see a situation in which AI was being applied to a practical problem within a company and a development team.

Observing the way my team worked together was also incredibly insightful. The completion of tasks and the documentation of work completed was done in a manner I had not seen before and was essentially another thing that my fellow intern and I had to learn along the way. This added another dimension to our work since we were not simply attempting to solve the problem we were assigned, but we were also figuring out how to relay our progress and findings in the most effective way possible.

2. RELATED WORKS

An article by Flynn, et al. (2024) discusses the prevalence of AI in the world of finance and Audit. The benefits of AI range from increased reliability and accuracy to the lowering of costs due to increased efficiency. Because of these benefits, an overwhelming majority of auditors and accounting firms plan on prioritizing the development of AI in their work in the coming years. Due to my work being in an accounting firm and using AI, I found the material in this article informative and valuable.

Jacobs, et al. (2022) provide insight into what Agile development is and the different terms and processes associated with it. The article details the process of backlog refinement, which essentially is the prioritization of work items based on functionality requirements compiled by critical stakeholders of an application. The article also goes into the importance of continuous integration and many other aspects of agile development that are widely utilized in software engineering as a whole. Some of these terms were things that I had to do during my internship, so this could help give insight into the agile development methods for readers.

3. PROJECT DESIGN

The issue that my team faced had to do with making a more efficient and integrated form recognizer with the company's systems. In addition, my fellow intern and I were tasked with researching potential AI solutions that could be implemented as a chatbot in the future.

The first thing we had to do, even before starting to work with AI, was become more familiar with the systems involved. These systems did not just include the technical stack we were working with, but also the procedures the company used in relation to agile development and work cycles. In addition, we had to gain access to project-specific softwares on our local machines. My fellow intern and I spent the first two and a half weeks just setting up our machines. We had to create several admin accounts on our laptops, and many of these took several business days to receive approval for, which delayed our set-up process tremendously.

In addition to initializing our admin access, we had to install several softwares to run the application. This included Visual Studio,

Visual Studio Code, SQL Server Management Server, Microsoft Azure Storage Explorer, and SQL Change Automation. Within Visual Studio, we had to download several extensions for the frontend code of the application to fully build and compile. We had to connect our Azurite blob service (used to store local instances of database objects in the application) to the Azure storage explorer through an extension in Visual Studio Code. We also needed to clone the repository of the application into our Visual Studios so we could seamlessly participate in version control to enable our changes to be reviewed and approved if matching company standards.

Additionally, the code for the application used the .NET framework and C#. I had no experience with either before this project, so much of my time went into familiarizing myself with the new technologies. All of this setup was complicated, and we got stuck many times along the way; but with help from each other and our supervisors, we were able to successfully run the application locally on our machines.

The next thing we had to do was familiarize ourselves with Azure boards, the platform the company used to keep track of sprints, work items, and big-picture project progress. These are all terms that describe agile development functions. Sprints are chunks of time, usually two weeks, in which a task is meant to be completed. Projects are broken down into sprints to make their progression more digestible and orderly. Work items, specific tasks assigned to developers in a given sprint, are usually given values for priority and difficulty. We were taught how to engage with this system by being given small bugs to fix in the application. We practiced being assigned work items, completing them, and

submitting a pull request—a review of our code by supervisors before it was integrated into the main application.

After this work was completed, we moved on to our project work with AI. We spoke to several other development teams at the company and they recommended that we use Microsoft Azure Document Intelligence to train AI models to extract data from various tax forms. Azure was already an industry-standard application because of the flexibility that came with its Neural models. We were able to train several models on W-9 forms and attempted to extend our use of the software to other tax form types, as well.

Last, I was able to do some high-level investigation into LLMs, more specifically Ollama for my project team. I found that one of the pros of Ollama is that it runs models locally, meaning it runs directly from a user's machine and not on a cloud. This ensures optimal security and on top of the customizability of Ollama, cemented the platform as a potential addition to the application.

4. RESULTS

After training several neural models, my team was able to implement the models into the application through the Microsoft Azure Form Recognizer API. This meant that we were able to successfully connect our model to the application and use it on user-submitted tax forms. The application utilized the trained AI model that we produced and the data was successfully extracted and stored in our databases.

We were only able to fully implement this process on W-9 forms as we did not have time to extend this to other models, though we trained more models to recognize other tax forms as well. By completing this task, we were able to open a door for the team we

were on to stop using their older OCR software, which they purchased from a third company license called ABBYY. This solution could potentially alleviate costs for the team since using a Microsoft Azure service would not incur extra expenses due to the already existing paid use of Azure services for other functions in the application. Moreover, by not using fully licensed third-party software and just utilizing a cloud service through an API, the form recognizer is also more controllable for the developers on the team since ABBYY had full control over their service.

5. CONCLUSION

My fellow intern and I were able to accomplish several tasks throughout our internship. We were able to assist in the implementation of a potential form reading AI solution, gaining familiarity with agile methodologies and professional task prioritization, and high-level research into an LLM that could be utilized in the future. Our work this past summer served to help the application of the team we were assigned to work more efficiently and be cost-effective. In the grand scheme of things, this allows for the firm to reallocate funds to other needs as well as for the application to be the best that it can be for partners and clients that use it.

My fellow intern and I were also able to gain a better understanding of the .NET framework and agile methodology. Through the work we did on this application, we expanded our understanding of model view controller frameworks (.NET is one) and became more familiar with the practices followed by top-tier software engineering companies and teams.

6. FUTURE WORK

The next steps regarding this project mainly involve an increase in scale. In regards to

the AI form recognizer, the immediate steps would be to train models to recognize every type of corporate tax listed in the application. This would allow the ABBYY software to be fully phased out of use in the whole application.

Additionally, a more long-term future step would be integrating Ollama into the application as an assistant chatbot for users. Essentially an entire front-to-end plan would have to be drafted with things such as user stories/requirements, functionality, UI/UX design, and testing all taken into account. This would be a fairly comprehensive widget and likely take several months to fully implement and integrate into the application.

7. ACKNOWLEDGMENTS

I would love to acknowledge my team, but particularly my fellow intern Amber, and my two managers Praveen and Aswathy. I learned a lot from every one of them and my journey would not have been the same without them!

REFERENCES

- Flynn, S., Mackenzie, R. and Moran, E., (contributors.) (2024) *AI in financial reporting and audit: Navigating the new era*. KPMG.
<https://kpmg.com/us/en/articles/2024/ai-in-financial-reporting-and-audit-navigating-the-new-era.html>
- Jacobs, M., Petersen, T., Kirsch, J., and Kaim, E. (contributors). (2022, November 28). *What is Agile Development? - Azure DevOps*. Learn.microsoft.com.
<https://learn.microsoft.com/en-us/devops/plan/what-is-agile-development>