Legna Software: Disrupting Forestry Ticketing with Machine Learning

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Matthew Gerace

Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Brianna Morrison, Department of Computer Science

Legna Software: Disrupting Forestry Ticketing with Machine Learning

CS4991 Capstone Report, 2023

Matthew Gerace
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
mjg5mm@virginia.edu

ABSTRACT
Legna Software is a Winston-Salem, North Carolina-based technology startup with the objective of digitalizing the forestry industry by leveraging a software-as-a-service model to provide wood mills with real-time financial and logistical information. One challenge is auditing mill-to-mill transactions. To address this, I leveraged C# as my programming language, Microsoft Azure as my cloud provider, and a non-SQL database to implement an end-to-end software application to make mill-to-mill receipt information instantly available on Legna Software's software-as-a-service model. This software application will eliminate paper tickets and the logistic errors and challenges that arise from this previous technology's delayed data transportation. Although the new model solves many problems, it can still be advanced through a more comprehensive machine-learning algorithm and stronger integration into the core system.

1. INTRODUCTION
I interned with Legna Software in the summer of 2022. Legna Software is a technology startup with a primary goal to digitalize and modernize the forestry industry by leveraging a software-as-a-service model to provide wood mills with real-time financial and logistical information.

One particular challenge is auditing mill-to-mill transactions. Mill-to-mill transactions arise from when a mill obtains excess or unnecessary trees, limbs, or branches while acquiring lumber. It requires a lot of resources to procure only necessary materials and trees, so transactions between mills with unwanted resources and mills desiring those resources happen frequently. As a result, wasting resources either disposing undesired materials or focusing solely on desired materials is avoided.

Through this selling process, the seller only receives feedback from the buyer after the lumber is obtained, delivered and then the driver returns because the information is not automated and instead travels handwritten on a receipt ticket. As a result, any issues regarding specifications or pricing come to fruition after two to four weeks. This challenge presented the opportunity for me to implement an end-to-end software application to make mill-to-mill receipt information instantly available on Legna Software's software-as-a-service model. Automating this flow of information will eliminate challenges and inefficiencies between the buyer and seller because now as soon as the buyer has an issue with the quantity, quality, or specifications of the material the seller is providing it is made aware to the seller. This new automated process, thus saves the seller money in delivering the correct materials and the buyer

money in resources to correct the seller's mistakes.

2. RELATED WORKS

Perweij, et. al. (2014) analyzes major research, work and applications of Optical Character Recognition (OCR). They describe the earliest uses of OCR: expanding telegraphy and creating reading devices for the blind. The modern version of OCR appeared in the middle 1940s and has been optimized to read forms, records and receipts. Understanding the transformation of OCR and its uses is directly applicable to my research in recording mill-to-mill receipt information because OCR is historically and practically one of the best solutions for digitally recording document information. Understanding these use cases provides insight for best practices and previously successful applications that will guide my utilization of this technology.

Subramani, et. al. (2020) surveys deep learning approaches for OCR and Document Understanding. They begin by discussing early rule-based algorithms that originated document understanding and introduce deep learning and natural language processing as a superior solution. They continue by highlighting document layout analysis, a subject matter essential for easily analyzing receipts and recording data. Document layout analysis is specifically applicable to my research in creating an end-to-end application for recorded printed information. Subramani et. al.'s research in data scarcity, alternative approaches, and science of information extraction is especially relevant for the data set curated for my research and its application.

3. PROCESS DESIGN

This section outlines the design process for the mill-to-mill transaction application.

3.1 Requirements

Both the client needs and system limitations influenced the application's design.

3.1.1 Client Needs

My introduction to this project began with an overall summary of the problem my work intends to solve. The problem being that mill-to-mill transactional information is not automated on the existing Legna Software platform.

3.1.2 System Limitations

This problem is due to not every mill being subscribed to Legna Software's platform, so in mill-to-mill transactions existing clients cannot get non-clients to use this platform. To address this issue, there is a need for an external application that can seamlessly input information into clients' Legna Software platform without being a customer and having the customer infrastructure.

3.2 Key Components

Using specifications and addressing challenges, led to design solutions.

3.2.1 Specifications

To create this external application, there were explicit specifications. The application needed to be able to connect with the main Legna Software platform which was hosted on Azure. As such, the application also needed to be hosted on Azure, which included being iteratively developed and saved with Azure DevOps, connected with Azure Data Storage, and using additional tools strictly from the Azure suite that were included with Legna Software's subscription. Furthermore, Legna Software utilized C# for many of their backend capabilities, so the application had to be coded in C# for connectivity and maintainability. Additionally, each mill utilizes separate receipt ticket structures, and because Legna Software's platform is applicable for any user

and their customers, it was required that any ticket regardless of structure could be accurately inputted into the system. Finally, my supervisor requested that the application not take up server space and costs and only utilizing resources when the data upload process was taking place

3.2.2 Challenges

Specifications regarding required technologies were challenging due to my lack of experience with Azure, C#, and Azure's Blob and NoSQL storage. However, the main challenge of individually developing this application was ensuring it was compatible with varying ticket structures. The second biggest challenge was ensuring the function application was only running when needed.

3.3.3 Solutions

Microsoft extensive documentation proved to be very useful in learning the required technologies and integrations between platforms. My experience with developing with VS Code and VS studio outside of the classroom also helped in having experience with compatible IDE's that enhanced integrations between my files and services. Microsoft Azure offered two key services that helped solved the two biggest challenges--The Azure Form Recognizer Studio and Event Grid Subscription.

The Azure Form Recognizer allowed me to leverage OCR and machine learning technologies to recognize unique ticket structures. I created machine learning models by taking photos of a sample of existing tickets and subsequently labeling desired fields on this service. Using these machine learning models with code enables the OCR values to be and exported into Azure Table Storage, so they can be integrated into the main platform.

Event Grid Subscription allowed me to ensure that the function app was only running on the server when it was needed to ensure resources were not wasted. This service enabled criteria for the application to service, which I set to be when a photo was taken and uploaded into Azure Blob Storage. Once the photo was uploaded, it was scanned with OCR, the necessary information was organized and recognized with the machine learning models and uploaded into Azure Table Storage. Finally, the image is moved into an archived folder ending the application's process

4. RESULTS

My application ultimately succeeded in recognizing ticket structure from photographs and uploading their information so it could be accessed by Legna Software's main platform while not using Legna Software's computing resources unnecessarily. The application leveraged 28 different neural network machine learning models I created from nearly 3,000 curated images. This ability to instantly upload information from any mill-to-mill transaction saved up to months of time in data transportation. Now instead of waiting for the ticket to make it all the way back to the mill after all the lumber has been cut and delivered to another mill, the information is available instantly.

I spent the last week of my learning experience taking the time to ensure my application was integrated into the system properly, well documented, and understood by my co-workers. It is still used by Legna Software in Winston-Salem as they continued to improve upon its user interface and accessibility. As Legna Software continues to grow and acquire more clients it will hopefully remain the backbone of clients' tracking mill-to-mill interactions as one of the first uses of machine learning

through form recognition in the forestry industry.

5. CONCLUSION

Creating an application to successfully read ticket receipt information has significant implications. Using OCR and machine learning in this context demonstrates practical uses of these technologies for efficient business practices. Furthermore, this application demonstrates that any industry is primed for technological disruption regardless of current technological influence and infrastructure and that this technological disruption can exist in the form of any technology. With the ability to correctly predict information pertaining to forestry metrics from a variety of tickets, there is room to not only improve this practice in this industry but expand it to different industries.

In the meantime, mills using Legna Software's products and future inspired technology, will save valuable from my application of these technologies. A simple upload of an image to read its contents using machine learning will save time, money and effort as information about products and transactions will be available instantly instead of months later. Optimizing an industry so influential yet technically behind helps our world advance through its improvement.

Furthermore, I improved as a software engineer. This project gave me great experience in learning new technologies, building a product from start to finish, and overcoming engineering challenges independently. I am proud to be a pioneer in revolutionizing the forestry ticket system industry by being one of, if not the first to introduce AI recognition.

6. FUTURE WORK

The use of machine learning and AI to recognize ticket structures and extract information can be expanded in the future to be more precise and applied to more industries. In the context of this specific application, more machine learning models can be created, more images could be used to train these models, and more combinations of these models could be tested to increase confidence in accuracy in prediction. While, this is the emergence of this technology in the forest industry, it has room to accelerate and make a substantial impact in other industries as well.

7. UVA EVALUATION

While my experiences and courses at the University of Virginia prepared me well foundationally to learn new ideas, practices and technologies quickly, there is still a lack of exposure to software engineering. The University of Virginia excels in teaching computer science, but with even more focus on practical software engineering, building end-to-end applications, and using technologies in the real world, my learning experience could have been enhanced. Specifically, it could have incorporated understanding and practicing AGILE development, deployment with DevOps tools and Cloud providers, and practice with modern languages and frameworks.

REFERENCES

Perwej, Yusuf & Hannan, Shaikh & Asif, Ali & Mane, Arjun. (2014). An Overview and Applications of Optical Character Recognition. International Journal of Advance Research In Science And Engineering. Vol. 3. Pages 261- 274.

Subramani, N., Matton, A., Greaves, M., & Lam, A. (2020). A Survey of Deep Learning Approaches for OCR and Document Understanding. ArXiv, abs/2011.13534