

Capstone Project Proposal



<Ashish Anand>

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	Problem: Currently, Billie's customers can upload multiple invoice documents in order to request financing; however, they are required to manually enter the relevant data associated with each invoice one at a time. This results in a slow and error prone process for customers to submit invoices on to our Factoring Platform. Goal: We are helping our customers to digitize and automate the invoice processing in order to improve their user experience, reduce time, and manual entries and errors on our factoring platform. How will ML/AI can provide value or help the business? ML/AI based OCR (Optical character recognition) can extract the text from the uploaded invoices both electronically generated and manually hand-written and process them without any need of customers to manually enter them in the system. Thus, make making it easy, less time consuming, and error-free.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	 On average our SMEs spent 5 minutes to manually fill the details of each invoices and they upload around 10-50 invoices each month. Since invoice number and other details are big and combination of alphabet, numbers, and special characters, they often made mistakes and request us to correct them later. On the other hand, due to lots of data-entry mistakes and inconsistencies, it slows down our internal audit work

	<p>and validation process of these invoices. It also increases burden on our customer support team.</p> <p>Automating the above process could potentially save 2-5 hours (could be more for some of our bigger customers) of our customers' time and improve their user experience on our factoring platform. Which can lead to higher satisfaction and saving for our customers.</p> <p>On the other hand, automation can also decrease the number of tickets handled by our customer support team. It can also help to fully automate the system, which can further reduce the operational cost.</p> <p>Standardizing and streamlining the process can also help the company to scale the business faster in different markets and countries.</p>
<p>Application of ML/AI</p> <p>What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?</p>	<p>There are two tasks in this project which can be best accomplish using below ML/AL model.</p> <ol style="list-style-type: none"> 1. Invoice image classification: It has been proved in many surveys by 2018 that AI/ML models are now performing better than human in image classifications. There are many free open sourced deep learning based multiclass models available which can be used to train to <u>precisely</u> identify invoice images from the other images. 2. Invoice text extraction: Once image is classified as invoice, next task is to extract the text from it, which can be <u>precisely</u> accomplished by deep learning-based attention OCR model. It can be thought of a of CRNN (Convolutional Recurrent Neural Networks) followed by an attention decoder. <p>Aim of the above explanation to explain what task ML/AI can accomplish. It's up to engineering team what model they select for these tasks.</p> <p>How does these two-model used to accomplish the below tasks:</p> <p>Invoice extraction</p>

	<ol style="list-style-type: none"> 1. Validate if the uploaded image is an invoice, if not, reject the invoice. (Image classification multiclass ML Model) 2. Detect and extract both electronically generated and hand-written text on the invoice images such as invoice number, invoice date, Tax %, Total invoice amount, issuers and receiver details. (Invoice text extraction) <p>Invoice validation:</p> <ol style="list-style-type: none"> 3. Validate the issuers and receivers are registered and supported by Billie's factoring platform. Whether it's a new or duplicate invoice. (Model will call API to validate it) <p>User validation:</p> <ol style="list-style-type: none"> 4. Populate the value on the UI after text extraction and validation. (Call API to pass the extracted value to User Interface) <p>Continuous ML Model learning:</p> <ol style="list-style-type: none"> 5. Once user corrects the data populated by the ML model on UI, this model will learn and improve the model with corrected data. (ML will consider user corrections as annotated training data) <p>Business outcome objective:</p> <ol style="list-style-type: none"> 1. Customer support saving 2. Improve NPS and customer satisfaction level 3. Reduce cost per invoice processing 4. Reduce invoice processing time 5. Reduce book-keeping cost
--	--

Success Metrics

<p>Success Metrics</p> <p>What business metrics will you apply to determine the success of</p>	<p>Business Metrics (OUTCOME):</p> <p>User adoption and engagement:</p> <ul style="list-style-type: none"> • % of customers started using this feature after pitch
---	--

your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.

- % MoM growth in user adaptation
- Conversion rate of automatic invoice process
- Average amount/invoice for manual vs automated processing

Task Success:

- Total time spent to upload and process invoice
- % times customers manually modified the value extracted by the model
- % reduction in customer support tickets

Retention:

- % of returned users (Churn rate)
- Average amount per invoice from returned customers

Happiness:

- % increase in NPS score
- % change in customer satisfaction score

Model Performance Metrics (OUTPUT):

- Precision, Recall, Accuracy of models
- Number of users and new users (Daily, weekly, Monthly)
- Number of users using automatic invoice extraction feature
- System performance: % time errors in upload and automatic invoice processing, API response time
- % of time ML model correctly extracted the invoice

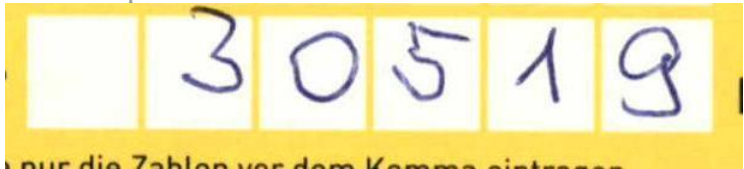
Baseline setting for above metrics:

1. For business metrics (Outcome), please consider industry benchmarks range (-5% to 5%) and percentage improvement over a defined period of time (Per week, month, or year)

	<ol style="list-style-type: none"> For model performance metrics (Output), trained model should have at least 80% of precision, recall, and accuracy. Since ML model should design to extract and display the value on the UI for user validation, activation function should avoid Step function which accept either “0” or “1” (Can be used Sigmoid, ReLU or any other similar function)
--	---

Data

<p>Data Acquisition</p> <p>Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?</p>	<p>Data Source:</p> <ol style="list-style-type: none"> There is no need to acquire data batch ongoing basis. Existing processed invoice images and manually entered value (real-time or production data) can be used to train the OCR ML model. Data acquired during prototype and usability testing. After production deployment, manually corrected data by real users to continuously improve the model. <p>Personal identifying Information:</p> <ol style="list-style-type: none"> Considering data privacy requirements Inform and educating users on how their data is processed and used by the company <p>Cost and availability of data:</p> <ol style="list-style-type: none"> Negligible cost and easily available All previously manually entered and processed invoices can be used as labeled data for training the OCR ML model. Sample size of 1000 receipts for each input class, such as all type of electronically printed image and hand-written invoices etc.
--	---

<p>Data Source</p> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>Biases:</p> <ol style="list-style-type: none"> 1. Over 90% of the processed invoices are electronically generated, which is <u>over-represented data</u> and can form biased 2. Different variety of handwriting styles which could result in poor recognition of some styles if model is not trained for majority of these style 3. Invoices printed in different languages can form biased if model is not trained equally for all language types 4. Different invoice image layouts. If model is trained for any specific layout such as horizontal or landscape, model can form a biased 5. If model is not trained with real-time or production data, model can break if it has to extract from real-world data. <p>How to avoid biases:</p> <ol style="list-style-type: none"> 1. Use real-world data to train model. 2. Use diverse set of data considering all used and edge cases 3. Use correctly labeled data 4. Use sufficient data to train mode. It is strongly recommended to use a <u>sample size of 2000</u> images of each class to <u>build and train a model</u>. 5. Use balanced data i.e. Dataset should incorporate all types of training data equally
<p>Choice of Data Labels</p> <p>What labels did you decide to add to your data? And why did you decide on these labels versus any other option?</p>	<p>For multiclass image classifier ML model, single label “invoice” will be sufficient because model just has to identify if it’s an invoice or not.</p> <p>For text extraction ML model, below labels for handwritten text extraction from invoice images: For example:</p>  <p>The image shows a yellow rectangular box containing the handwritten numbers '30519' in blue ink. Below the box, there is a line of text in German: 'nur die Zahlen von dem Konto eintragen'.</p>

	<p>What is the state of the image?</p> <p>Broken or Not Loading</p> <p>Showing but Illegible</p> <p>Legible</p> <p>For text extraction ML model, for electronic generated bills, only below labels should be added for fast image transcription:</p> <ol style="list-style-type: none"> 1. Invoice number 2. Invoice date 3. Due date 4. Tax % 5. Total invoice amount 6. Issuers details <p>Labeling only relevant accounting terms could help the machine to learn faster and efficiently. Further, in this functionality, model has to learn and extract only above labels (accounting terms) and its corresponding details. Additionally, it will also reduce computational burden on the nodes.</p> <p>Since there are few predefined labels, training model for other languages would also be easier. This could particularly help the company to scale the product internationally.</p>
--	--

Model

<p>Model Building</p> <p>How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?</p>	<p>ML Model recommendation:</p> <p>We should consider using customized model which can leverage existing open source libraries, because of the following reasons:</p> <ol style="list-style-type: none"> 1. We need to deploy two ML models. One has to tailored to perform generic task like identifying invoices and other has to tailored to perform specific task like extracting accounting related text from invoices.
--	--

	<ol style="list-style-type: none"> 2. Outsourced model such as Google AutoML can't support all the use cases we have for this project. Further, there are plenty of open source available for image recognition and OCR, which can easily be customized and trained for all the use cases of this project. 3. Accounting and transaction data are confidential, it's risky to share this with third-party.
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	Evaluating results: As mentioned above, we need to deploy two ML models: <ol style="list-style-type: none"> 1. Image Identification (Static ML Model) 2. Text Extraction (Dynamic ML Model) <p>To measure the success of first model - Image Identification, more emphasis is to gain high recall, avoiding false negative cases as much as possible and even at the cost of low threshold. We should aim to get recall at least 90%.</p> <p>To measure success of second model - Text Extraction, we should aim to get high precision at least 0.8 or 80%. Accuracy of extraction could be compromised initially because of two reasons:</p> <ol style="list-style-type: none"> 1. Users will have the opportunity to correct the extracted data on the screen before he or she submit it for processing. 2. Since it's a dynamic model, model will learn in real-time with every time use corrects the data and submit.

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches	The features in the MVP will be limited as following: <ol style="list-style-type: none"> 1. Option to users to upload multiple invoices, both handwritten and electronically generated (Refer
--	--

of your product.	<p>appendix: Wireframe Number 1: Upload bills)</p> <ol style="list-style-type: none"> 2. Validate invoices images, identify duplicates invoice and remove invalid images (Refer appendix: Wireframe Number 2: Uploaded invoices) 3. User notification once invoice is extracted and its values is populated on the screen 4. Display the invoice uploaded, Option to use to navigate to all uploaded invoices, and further, review, correct the value extracted by the model 5. Option to retrieve the invoice uploaded in past session and edit it or/and submit it for further processing ((Wireframe for feature 3,4, & 5:
<p>Use Cases</p> <p>What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?</p>	<p>Use cases are for B2B clients (SMEs).</p> <p>Personas here would be:</p> <ol style="list-style-type: none"> 1. Head of Finance – Early stage Start-ups (1 to 50 Staffs, 0-10 invoices) 2. Procurement or Finance department staff, Growth or medium sized company (50 - 200 staffs, 10-50+ invoices) <p>Epic level use cases:</p> <ol style="list-style-type: none"> 1. User should be able to upload his or her invoice(s) 2. Once uploaded, user should be able to see appropriate extracted text auto-filled for the invoice(s) 3. User should have option to navigate, review, and update all the invoices he or she uploaded 4. User should be able to edit the invoices details unless he or she submitted it to process
<p>Roll-out</p> <p>How will this be adopted? What does the go-to-market plan look like?</p>	<p>Pre-launch planning:</p> <ol style="list-style-type: none"> 1. Testing and pilot phase: Initially tested by engineering team and other internal stakeholders for functional, performance, and compatibility testing. After that, prototype will be shared with selective customers to test with their real data. Based on self-observation and clients' feedback, prototype will be re-

	<p>iterated and send to these selective customers to validate it again.</p> <p>Launch planning:</p> <ol style="list-style-type: none"> 1. Set-up a plan with marketing team to promote OCR features and its benefits. Use the metrics of final test results for marketing. 2. Use customers as earned influencers to promote the feature and the product. They will happily promote, since their problems are solved, and they are satisfied. 3. Run A/B testing with this new feature. Based on both qualitative and quantitative data, decide to roll-out or not to all users 4. Reach-out to all the existing customers and update them on this new feature and launch date. 5. Set-up after launch support to support customer using OCR feature <p>Post-launch planning:</p> <ol style="list-style-type: none"> 1. Listen to the feedbacks and fix them in future iterations 2. Add new features that support different language, market and countries 3. Implement features that improve user experiences 4. Keep track of financial(P&L) and non-financial KPIs.
--	--

Post-MVP-Deployment

<p>Designing for Longevity</p> <p>How might you improve your product in the long-term? How might real-world data be different from the training data? How will</p>	<p>Important future product improvements:</p> <p>MVP offers some basic features such as browse and upload the file and extract the text and autofill the details on the webpage.</p>
---	---

your product learn from new data?
How might you employ A/B testing
to improve your product?

1. Add different language and handwriting support: Model has to deploy and support to many geographic locations, where it has to recognize different languages, handwritings etc.
2. Mobile version or app can use phone camera API to scan the image. This way, use will not be having hassle to scan and download to memory device before they upload it to the platform.
3. Option for the users to upload files in different formats. Platform currently support only pdf format, but it can support other common formats. such as jpg, png etc.

Training data vs real-world data and how product will learn from new data:

In this project, we are using real world data to train our models. However, for handwriting recognition could be tricky and it would be challenging for the model to predict it correctly because there are lots of variation in it and it's practically impossible to train the model with all handwritings.

To learn product from new data, I have proposed to use human-in-loop approach without hiring any additional dedicated human annotator. When text will be extracted by the uploaded invoice of the user, it will ask the user to review and submit. Once user review and update the value, Machine will learn corresponding data label and after reaching such threshold, it will start predicting in similar way. Same as when we move the certain emails in the spam, algorithm learn it and set a rule automatically to do so in the future.

**How A/B testing will be implemented:
Steps:**

1. Review the customer feedbacks and metrics to identify issues and area of improvements.
For example: Suppose we found that, most of the users are not updating the auto-filled values but he or she is deleting in and typing it from scratch. This possibly means that we need to adjust

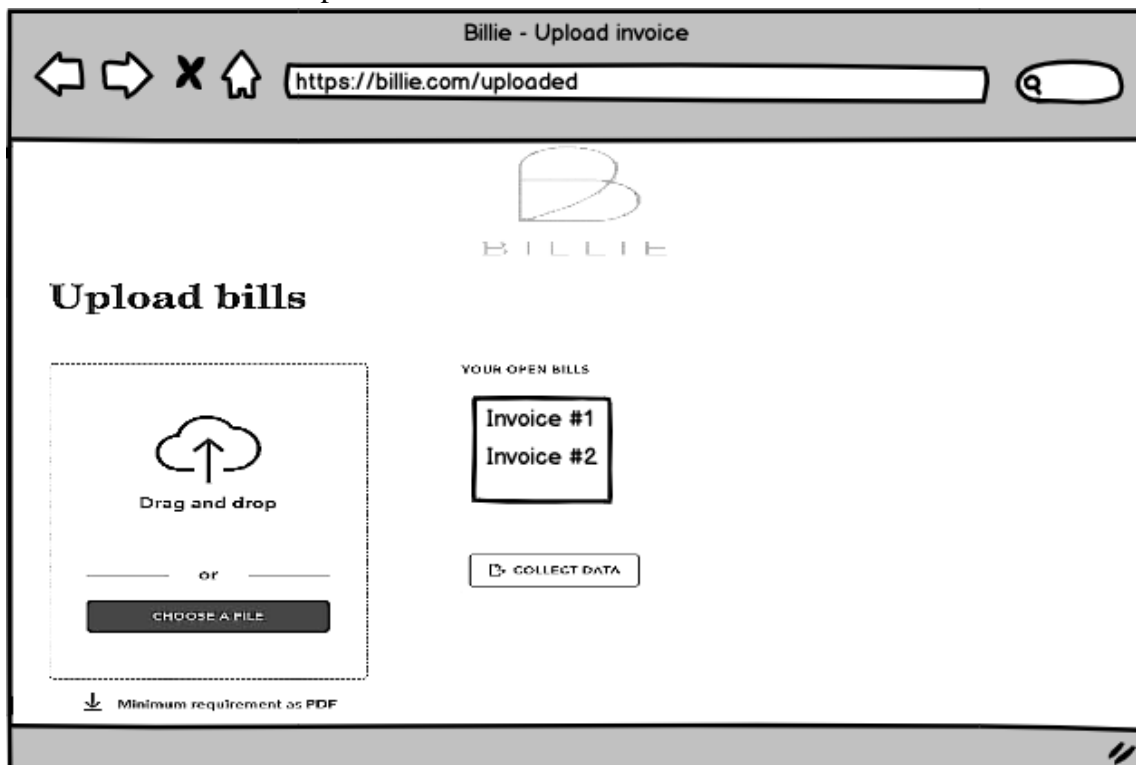
	<p>precision and confidence threshold values of the text extraction model</p> <ol style="list-style-type: none"> 2. Design null (Ho) and alternate hypothesis (Ha) statements with new precision and confidence threshold values of the challenger model 3. Look into the metrics to verify the changes. Choose an unbiased sample and run the experiment to observe if this changes the users' behavior. 4. Based on that outcome, if needed, replace the controlled model by the challenger model.
<p>Monitor Bias</p> <p>How do you plan to monitor or mitigate unwanted bias in your model?</p>	<p>In this project, Annotation bias is less likely to be developed, since model is dynamic and most of the annotation would be done by the real user with the real data.</p> <p>However, it's more likely that model would develop unwanted data bias.</p> <p>Since OCR model is most likely to extract information correctly from electronically generated invoices. Most of the corrections (also human-in-loop data annotation) would be done for the handwritten invoices. This will eventually generate most of the training data from manual invoices. Further, this can form model staleness for computer generated invoices i.e. decrease in the model performance over the time.</p> <p>To mitigate this, team should consider continuously optimization and training of the model with balance data i.e. training data should have equal number of electronically generated invoices and hand-written invoices. It is also recommended that this is done continuously and frequently since it's a customized and dynamic ML model.</p> <p>Additionally, It is also recommended that model are regularly updated with the latest open source libraries to ensure even incompatible handwritings are understood reducing the hassle of end users and also reducing the bias introduced by handwriting and data annotation.</p>

Appendix: Wireframes

Wireframe Number 1: Upload bills



Wireframe Number 2: Uploaded invoices



Wireframe Number 3: Review / Update Invoices

Billie - Enter or Update details

https://billie.com/upload/edit

BILLIE

Enter your billing information

Please complete and confirm your billing information

Invoice recipient

primaholding GmbH

Bill number

R0123

Gross invoice amount

1.020,99 €

VAT rate

19%

Date of invoice

9 Oct 2019

Due Date

22 Nov 2019

NEXT TO FINANCE

Minimum requirements checked?