**Capstone Project**

**Documentation**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing, and delivering the Capstone project.



# Problem statement

## What is the problem being investigated?

Possible Me & Co. is a start-up in the mobile application development space. The product owner had no time to read the customer reviews concerning the mobile application. Because most of the people working at the start-up were busy freelancers, the owner proposed that the data scientist create a micro-application to help automate the process of evaluating customer sentiment on a product, to prioritize tasks for the IT team. I, the data scientist has been tasked with resolving this issue for both the short and long term.

## Why is this problem valuable to address?

* This will enable the team to have a ground truth to evaluate whether or not customers are happy with the product without having to read through every single review on every product in the portfolio.
* Addressing this problem with big data tools allows the process to be integrated into the workflow pipeline, allowing teams to immediately see the sentiments of customer reviews in real time.
* Having sentiment analysis as part of the data pipeline enables the team to get a visualization metric that helps visualize the effect of feature deliverables quickly. This means a faster mean time to error (MTE) than before.

## What is the current state?

* The manager has a barrage of customer reviews for 20 different mobile applications he is managing at once.
* The communications between staff at Possible Me are not as productive as other start-ups. Outdated workflow technologies are still used. Even though team members have proposed that the manager use better project management software and methodologies, the response is unconvincing.
* As a result, once an application feature is up and launched as a beta test, the next iteration of Kanban meetings is only decided once the blooming project hype has already died down. Hence most of the users go to other competitors for better substitute products.
* Customer complaints are not attended to as too much attention is focussed on applications that are already doing well. As a result, prioritization of fixing declining software relies on understanding customer sentiment of an app at a glance.

# Industry/ domain

## What is the industry/ domain?

The domain being discussed is the software development domain, more specifically mobile application development for Android and iOS. Applications can be made for mobile or for the web. Usually a start-up like Possible Me will want to do both.

## What is the current state of this industry?

The mobile application market was valued at 154.05 million USD and is forecast to grow at a 11.5% compound annual growth rate (CAGR) from 2020 to 2027. The growth is primarily due to the rise in Internet usage by developing countries such as China, India and Brazil.

Also, with new technologies like 5G, ecommerce, IoT, machine learning, AR and VR, gamers and shoppers a like have been purchasing more premium applications from the App Store or Google Play Store.

While the surge in the ability of technology to interact with the consumer is a good thing for sales, this also means that Cost of Goods Sold (COGS) tends to increase. Even while sales may increase in parallel with innovation, IT teams have become more pressured to speed up cycle time delivery to keep up with ever changing demands. This is why IT teams are now adopting Agile approaches to speed up software delivery.

## What is the overall industry value-chain?

Ever since the backlash of the 1980s towards heavyweight processes to software delivery started to gain steam, a new movement called the Agile Manifesto began to take root. Frustration was looming over Waterfall methodologies that utilized a single phase Software Development Life Cycle (SDLC) from Planning to Maintenance without revisiting the planning phase. So, when the project was finished, the product became obsolete, and the competition moved on.

Modern mobile development uses one of these Agile methodologies. These include scrum, Kanban, extreme programming (XP), DevOps and others. As such the Agile value-chain is still similar as the Waterfall model, going through the 7 stages: Planning, Requirements Analysis, Design, Development, Testing, Implementation and Integration, Operations and Maintenance. However, Agile is an interactive approach.

For example, in Scrum, the IT team keeps a backlog of ideal corrections to the list of functional requirements. The product owner must then manage the scope creep at the next meeting or stand up. This process is repeated constantly in “time-boxes” and as such the value-chain in software development tends to be iterative and fast.

## What are the key concepts in the industry?

* SDLC: Software Development Life Cycle
* The Agile methodology: popular reference to the Manifesto of Agile Software Development
* Concurrency: This is how process requests are interleaved. For example, in an Online Transactional Processing (OLTP) system, transactions tend to be isolated if there is little tolerance for error. However, Online Analytical Processing (OLAP) systems tend to allow more interleaving trading off security for efficiency.
* Containerization: This helps us to build scalable applications that are operating system agnostic. This means that we can launch an app in multiple environments as the code is not dependent on any infrastructure. This is great for scalability as multiple code versions of a same product need not be replicated.
* Responsive Web Design: With the rise of modern web development frameworks like Angular and React, responsive web design is now more important than ever to capture market share. Techniques in this space can be invaluable as it provides the product competitive advantage.
* Processing: In cloud mobile computing, we often talk about batch processing and stream processing. Stream processing is good for real time use cases, such as image recognition of invoices giving instant feedback to the user. Batch processing is a lot cheaper in most cases, as it processes its workload in batches. Often batch processing is more efficient as it does not need a lot of resource overhead and can be scheduled as when needed.
* Workflow: When software people talk of workflow, they are often referring to CI/CD processes. This can involve software like Jira,

# Stakeholders

## Who are the stakeholders?

|  |  |
| --- | --- |
| **Stakeholders** | **Concerns** |
| Customers | *Problem:* frustrated that their complaints in the Google Play App Store are not being responded to.  *Expectations:* thinks the problem should be fixed in a week |
| Manager | *Problem:* flustered with multiple project deadlines. Every time he wants to report budget figures to the financial manager, he is always put on hold as his team take a long time to decide.  *Expectations:* |
| Product Owner | *Problem:* ignores scope creep of projects because of time constraints, even if some scope creep may improve the product and raise revenues for the period. |
| Requirements Analysts |  |
| UI/UX Designers |  |
| Web developers |  |
| Java Developers |  |
| Data Engineers |  |
| Financial Manager |  |
| Marketing Manager |  |
| IT Project Manager |  |

# Business question

## What is the main business question?

* Question: How to increase delivery cycle speeds by prioritizing applications with the lowest perceived value from customers?
  + Answering this question in a data-driven way will help the IT team focus on high-priority applications. This means that applications that are already doing well can be focussed on later.
  + The analysis of sentiments of customers will aid this process of discovering the appropriate solution.

## What is the required accuracy?

* Since the proposed data solution would likely be a machine learning algorithm that tries to predict the sentiment of customers, we want the solution to be able to be reproducible and updatable throughout its lifetime.
* It is also important that the data pipeline can bring in fresh data every time the code is run. This can be in the form of a webhook or API call.
* A baseline accuracy of 80-85% (Paul Barba, 2019) would be reasonable for sentiment analysis, although surpassing that baseline would be ideal.
* ­­For this use case, false negatives would be more important than false positives. In other words,

­­­

# Data question

* What is the data question that needs to be answered?
* What is the data required to answer the question?

# Data

* Where was the data sourced?
* What is the volume and attributes of the data?
* How reliable is the data?
* What is the quality of the raw data?
* How was this data generated?
* Is this data available on an ongoing basis?

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
* What are the highlights of the Exploratory Data Analysis (EDA)?
* Is the pipeline reusable? (for example, to process future data?)
* What are the intermediary data structures used (if any)?

## Modelling

* What are the main features used?
* Did you find any interesting interactions between features?
* Is there a subset of features that would get a significant portion of your final performance? Which features?
* How did you select features?
* What feature engineering techniques used?
* What are the models used?
* How long does it take to train your model?
* What are the tools used? (cloud platform, for example)
* What are the model performance metrics?
* Which model was selected?

## Outcomes

* What are the main findings and conclusions of the data science process?

## Implementation

* What are the considerations for implementing the model in production?

# Data answer

* Was the data question answered satisfactorily?
* What is the confidence level in the data answer?

# Business answer

* Was the business question answered satisfactorily?
* What is the confidence level in the business answer?

# Response to stakeholders

* What are the overall message and recommendations to the stakeholders?

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

# References

* <https://www.grandviewresearch.com/industry-analysis/mobile-application-market>
* <https://agilemanifesto.org/>
* <https://www.investopedia.com/terms/v/valuechain.asp>
* <https://techgeek365.com/5-software-development-concepts-to-improve-your-programming/>
* <https://www.lexalytics.com/lexablog/sentiment-accuracy-baseline-testing>