**Capstone Project** **Document**

DSIA – Data Science and AI Course

Part-time – 2020-06-23

# Len Kwok

# Process overview

A picture containing green, sign, parked, device

Description automatically generated

# Problem statement

* What is the problem or the opportunity that the project is investigating?

To predict if Auto ML can help data scientists in their work.

* Why is this problem valuable to address?

Because in 2013 alone, 23 billion dollars is wasted on big data investment by companies.

* What is the current state (e.g. unsatisfied customers, lost revenue)?

The amount today is even more than that in 2013 as data created keeps on increasing.

The estimated annual expenditure with CVD hospitalisations in Australia is

$8.14 billion (improving efficiency by 1% means saving $81.4 million). With more complex algorithms, data scientists need a more efficient way of working.

* What is the desired state?

Data scientists can answer business and data questions at scale, with good accuracy and get ready for production handover.

* Has this problem been addressed by other research projects? What were the outcomes?

Perhaps. While auto ML is not new, the ability to create the entire data pipeline and quickly get models into production is still a novelty.

# Industry/ domain

* What is the industry/ domain?

Data science and all industries in general.

* What is the current state of this industry? (e.g. challenges from startups)

Increasing amounts of data created, complexity of data science, lack of manpower, wastage of big data budges.

* What is the overall industry value-chain?

CEOs > Data Scientists > to anyone in general

* What are the key concepts in the industry?

The objective is to unlock the value in data so as to improve business metrics.

* Is the project relevant to other industries?

Yes, to all industries.

# Stakeholders

* Who are the stakeholders? (be as specific as possible)

CEOs, CFOs, marketing department, data scientists, data engineers, data analysts, dev ops, solution architects.

* Why do they care about this problem?

Covid 19 has made past records redundant and we all need new data to predict outcomes and make better decisions. It is expensive to procure new data, not to mention time consuming as well.

* What are the stakeholders’ expectations?

Some stakeholders in the healthcare sector have conflicting interests. Technical staff have data metrics in mind while business leaders care mostly about financial KPIs. With Auto ML, data scientists can automate the tedious aspects and concentrate on domain expertise, so as to solve the business questions at hand.

# Business question

* What is the main business question that needs to be answered?

In 2013, 23 billion dollars was wasted on big data investment. It is now 2020. With even more data created and the presence of Covid-19, the amount of money wasted is even more.

* What is the business value of answering this question? (quantify value and make necessary assumptions)

Companies might be able to save 23 billion dollars at the minimum if they have coherent data strategies.

* What is the required accuracy? What are the implications of false positives or false negatives?

The best accuracy as we can get from python coding versus that obtained by Auto ML. This is a straightforward binary classification problem that exists in real life.

# Data question

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

Is Auto ML more accurate than us in predicting whether bank credit card customers are ‘risky’ or not?

* What is the data required to answer the question?

Dataset contained 5,000 entries of German credit card customers.

# Data

* Where was the data sourced?

The data was provided by IBM.

* What is the volume and attributes of the data?

5,000 entries. 20 features.

* How reliable is the data?

Extremely reliable.

* What is the quality of the raw data?

Very good.

* How was this data generated?

IBM made this available as part of Auto ML tutorial.

* Is this data available on an ongoing basis?

Yes. But only accessible if a user signs up for IBM Cloud Pak.

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?

We did data wrangling on our own at first. Auto ML is able to perform the same process.

* What are the highlights of the Exploratory Data Analysis (EDA)?

Auto ML can not only perform PCA analysis, it is able to perform feature engineering including creating new features. The slick UI enables us to perform EDAs easily and show more graphs. We humans, can do it, but Auto ML is much faster.

* Is the pipeline reusable? (for example, to process future data?)

Yes. Totally.

* What are the intermediary data structures used (if any)?

Not that we are aware of. It must be mentioned that auto ML used some of the existing features to create new ones.

## Modelling

* What are the main features used?

Checking Status, Loan Duration, Credit History, Existing Savings, Employment Duration, Instalment Percent, Sex,

Current Residence Duration, Age, Instalment Plans, Housing,

Existing Credits Count, Job, Dependents, Telephone, Foreign Worker

* Is there a subset of features that would get a significant portion of your final performance? Which features?

Loan Duration, Age, Job.

* How did you select features?

Bagging.

* What are the models used?

We used decision trees, random forests, bagging and boosting, all with cross validation. Auto ML used at least 8 algorithms for classification including logistic regression.

* How long does it take to train your model?

A few minutes, with this data set. It must be said that auto ML did it faster with multiple pipelines simultaneously.

* What are the tools used? (cloud platform, for example)

Google Colab, with GPU. Jupyter notebook. IBM Cloud Pak portal

* What are the model performance metrics?

We focused on accuracy as the main metric. It must be noted that Auto ML was able to generate more metrics.

* Which model was selected?

For us, bagging. For Auto ML, gradient boosting. Because multiple pipelines were generated, there is an option to pick and choose which one we like.

## Outcomes

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

Instead of making data scientists obsolete, auto ML can make data sceients more efficient. The wealth of results can be examined further to gain more insights into customer behaviour for example. Hence, companies have more confidence when they invest in big data. More than ever, data scientists can meet the present and future demands better than before.

## Implementation

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

Auto ML makes it very easy. It can generate forms to capture new data singly or by batches using APIs. Once put into production, it can monitor for drifts in data. It all depends on the cost of investment in auto ML. If put to good use, it can certainly be cheaper than $23 billion in 2013.

# Data answer

* Was the data question answered satisfactorily?

Yes, we have identified ways that auto ML can help to make data scientists perform their jobs better with great accuracy.

* What is the confidence level in the data answer?

We obtained 81% accuracy while auto ML obtained 80.7%.

# Business answer

* Was the business question answered satisfactorily?

Yes. Data scientists are now able to use automated tools at their disposal to help companies achieve cost savings in big data investment,

* What is the confidence level in the business answer?

Same as above.

# Response to stakeholders

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

With proper use of auto ML, there will be less conflict to all stakeholders in terms of achieving data and business metrics.

# End-to-end solution

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

The entire data pipeline including deployment for production is saved onto IBM Cloud Pak portal for easy retrieval and duplication if need be. The portal generates a Jupyter notebook for editing.

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

Available at: <https://github.com/MirunaSuresh/Jun23cohortDataScienceAndAI/tree/master/Len/Capstone%20Project>

* What are the resources used in the project? (libraries, algorithms, etc)

Available at: <https://github.com/MirunaSuresh/Jun23cohortDataScienceAndAI/tree/master/Len/Capstone%20Project>