SkillsFuture Assignment

Mark Lim

Section 1

Objective: To discover insights and trends in data science from the Kaggle data science survey from 2017 - 2021

For management, charts should usually be intuitive and quick to understand. In this case, I have used simple bar charts along with narratives for better understanding

Background

Trends and Insights Analysis from Kaggle Data Science Survey

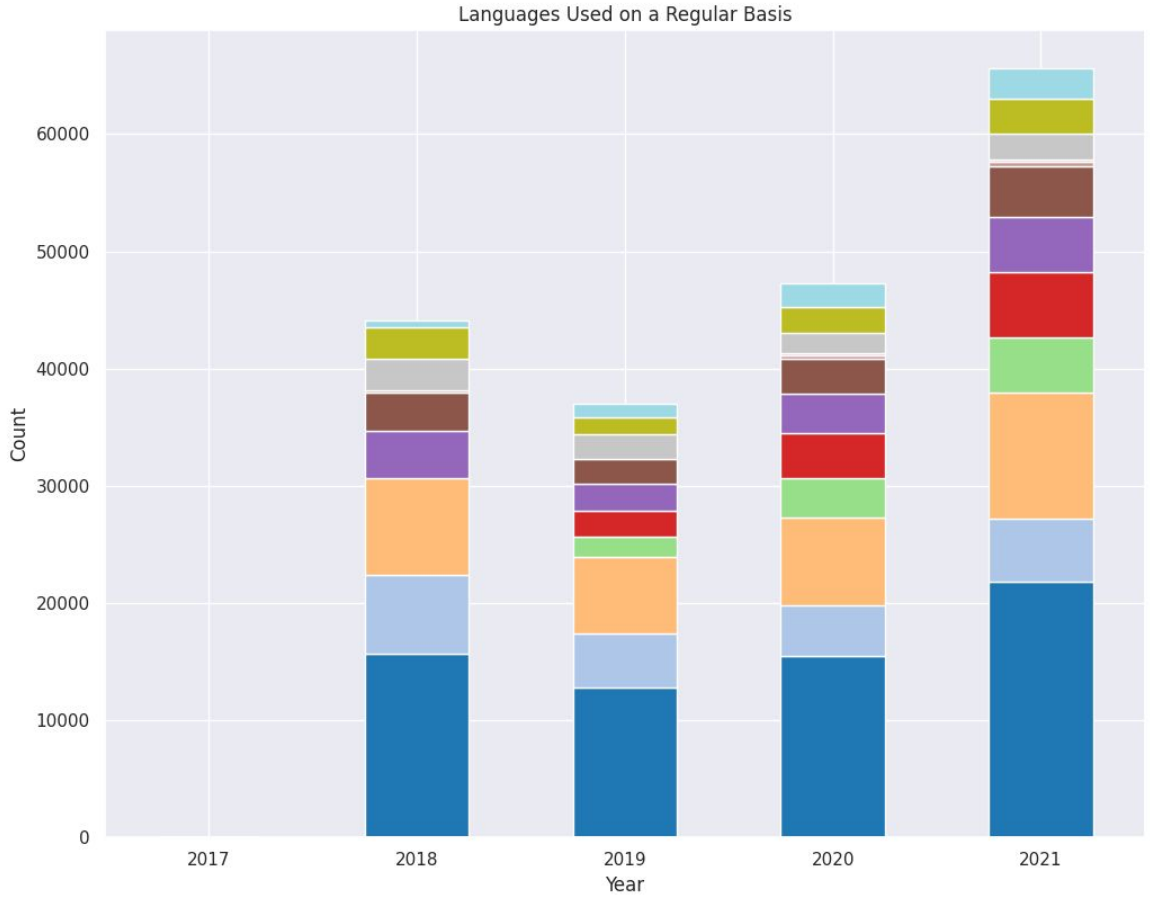
About the data
The Kaggle Survey Dataset is a dataset that comprises survey answers covering different aspects, aiming to capture various information about Kaggle users. The dataset spans from 2017-2021.

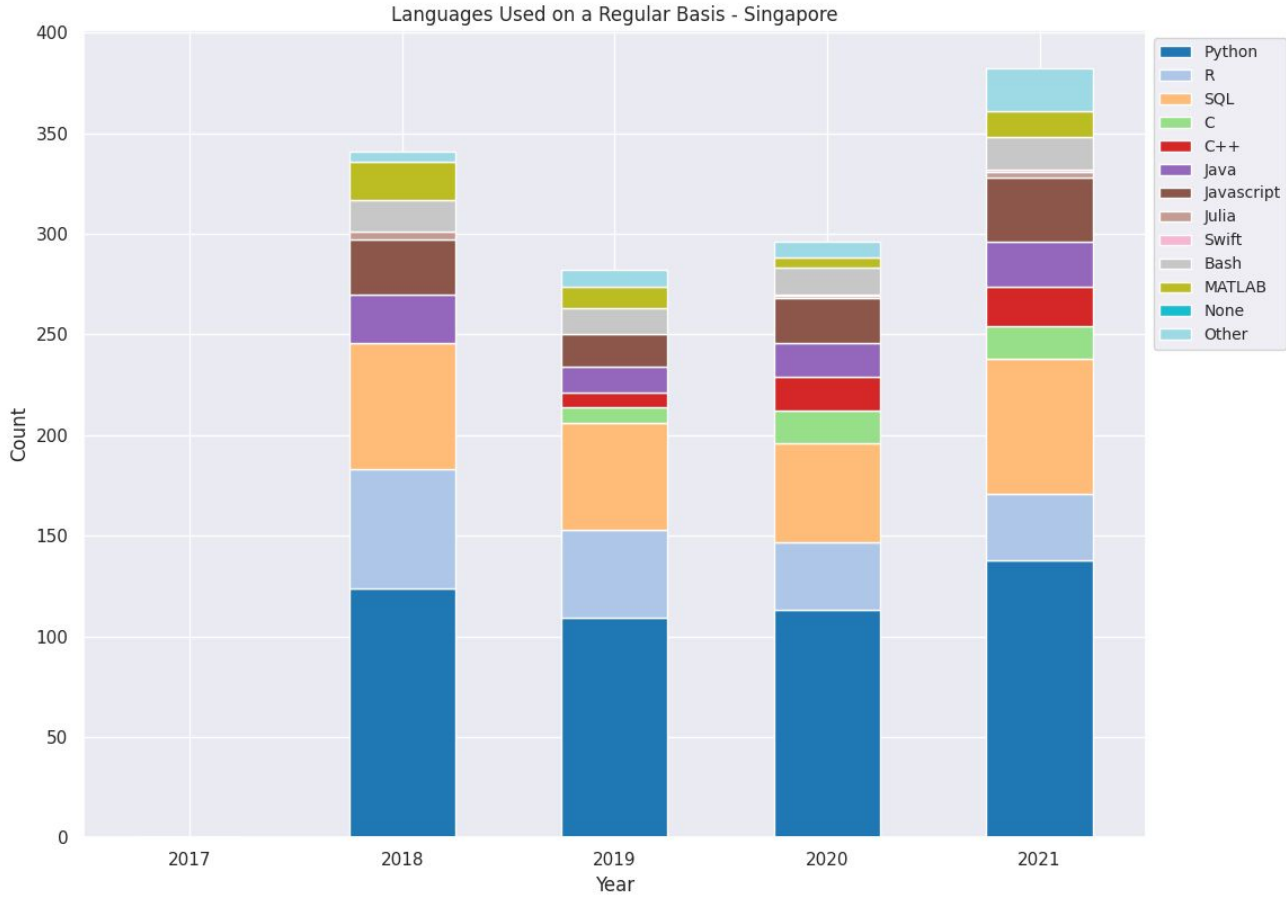
3 main areas of investigation

- Technical Knowledge
- Domain Knowledge
- Education and Experience

By investigating these areas, we can identify market demand and gaps which can be filled

Technical Knowledge - Programming Languages



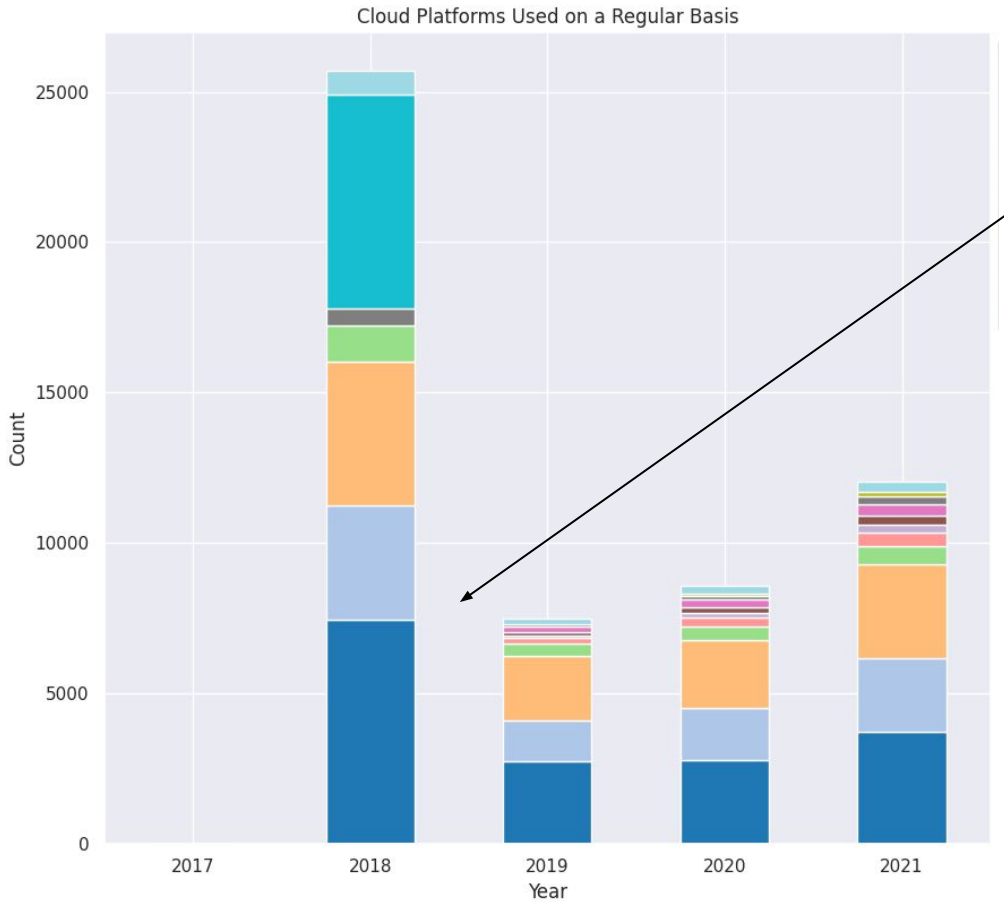


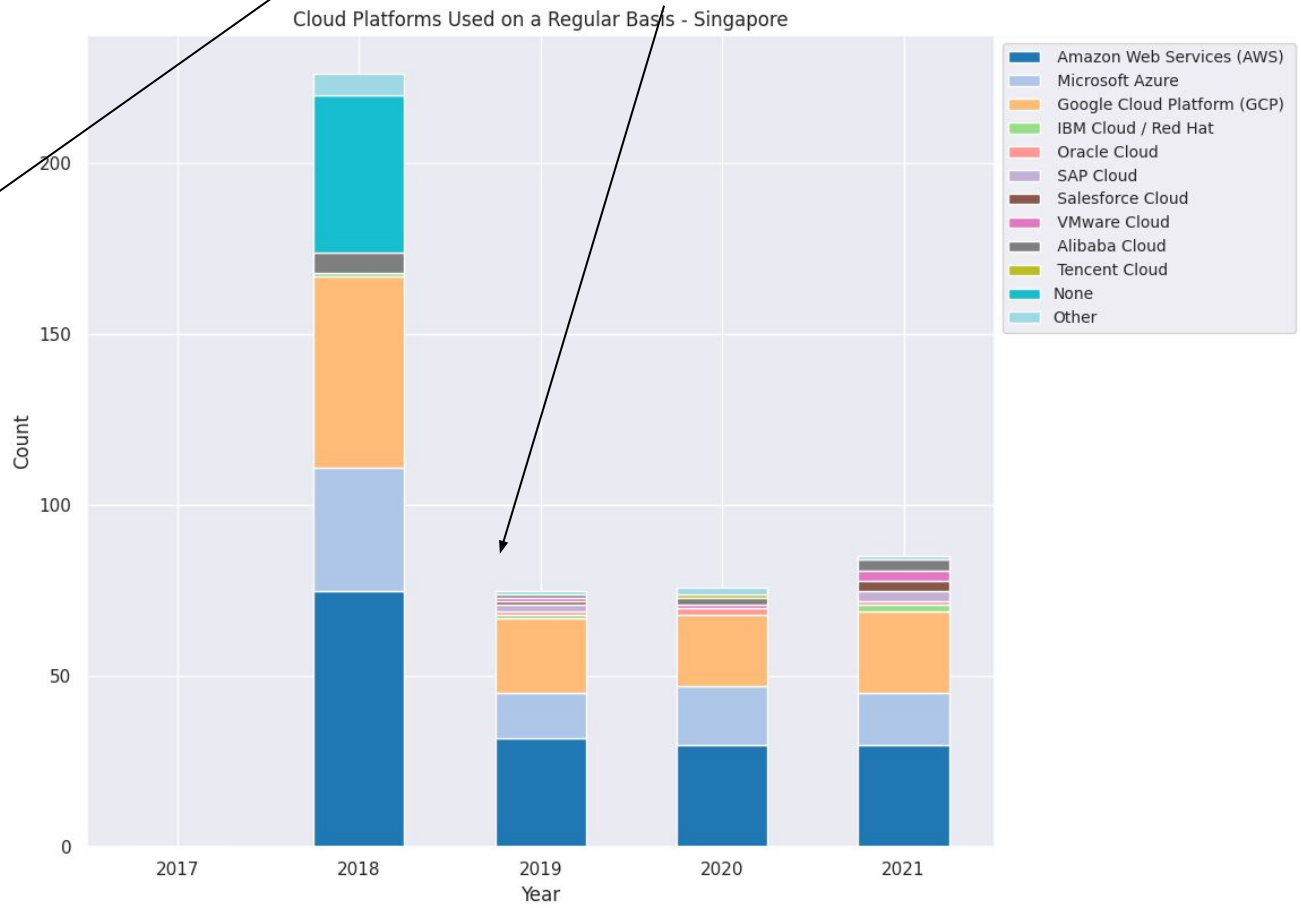
Insights:

- Python remains the predominantly used language, followed by R and SQL, from 2018 - 2021.
- This trend is observed both globally and in the Singapore context.
- Knowledge in these 3 languages remains key in the realm of data science. Were SkillsFuture to encourage individuals to enter this field, efforts should be focused on increasing literacy in these 3 languages.

Technical Knowledge - Cloud Platforms

Notice that the volume of the responses from 2018 drop significantly to 2019. Thus we consider only responses from 2019 to 2021, where response volume is more consistent

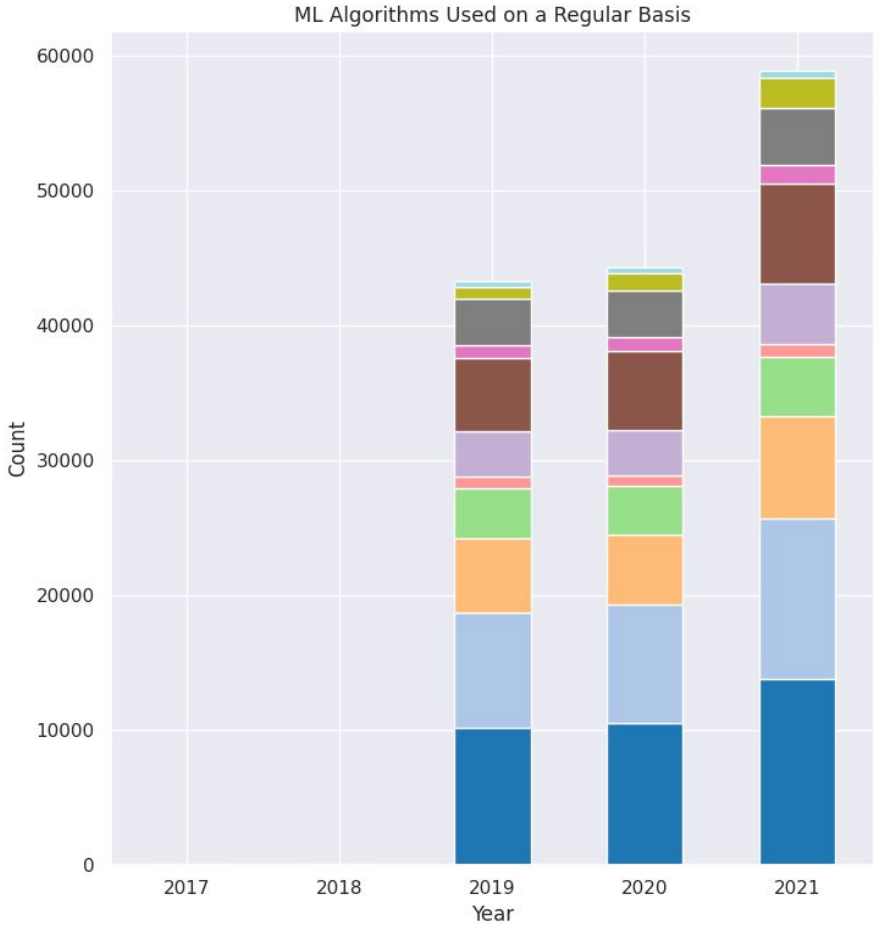


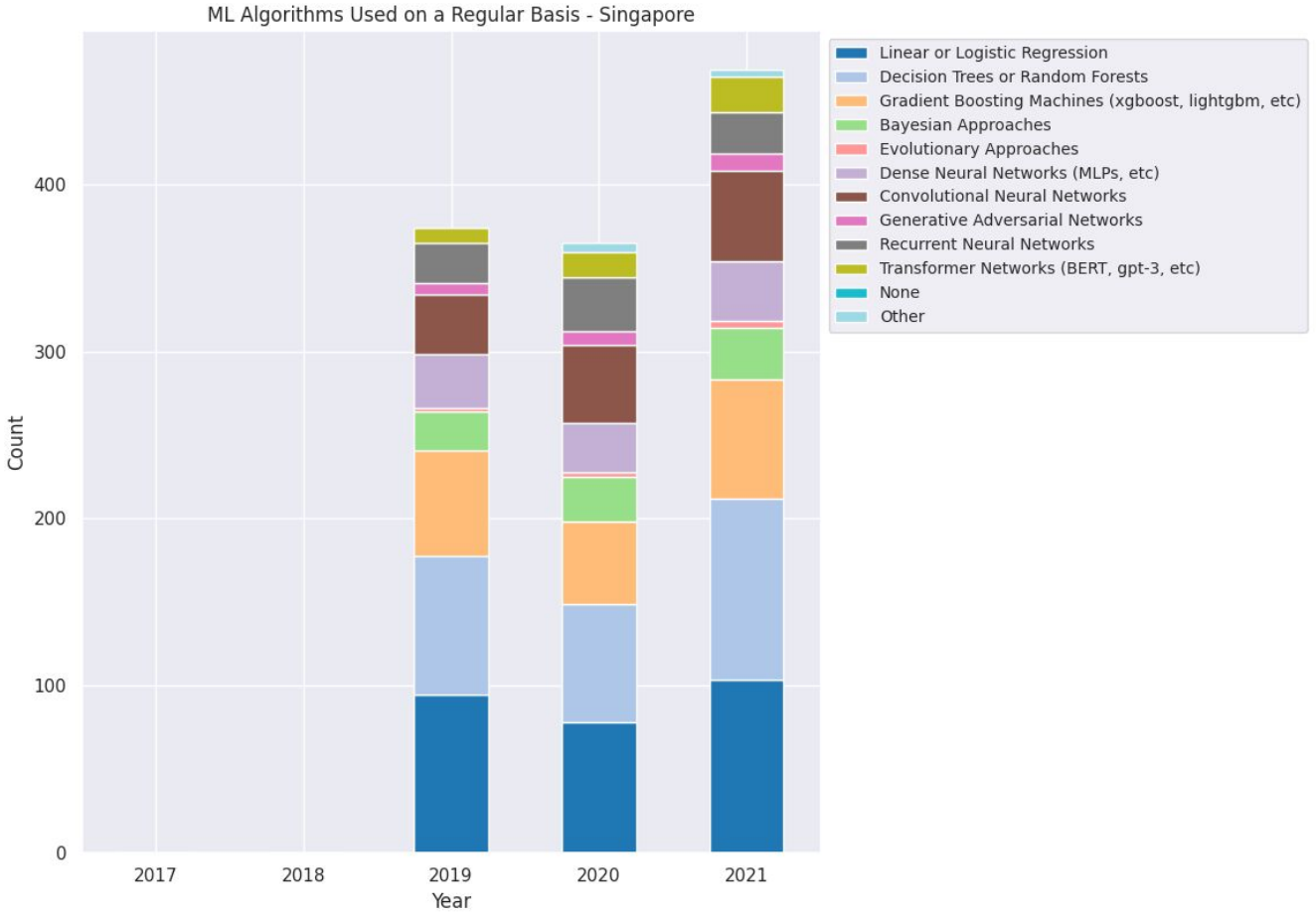


Insights:

- AWS, Azure and GCP are the 3 most used cloud platforms for data analytics
- Considering that these enterprise level tools are used regularly in the field and it is difficult for an individual to gain access to these tools without a subscription, SkillsFuture can consider offering courses to teach these tools that come with student APIs, for individuals to gain some hands on experience with these tools

Domain Knowledge - ML Algorithms

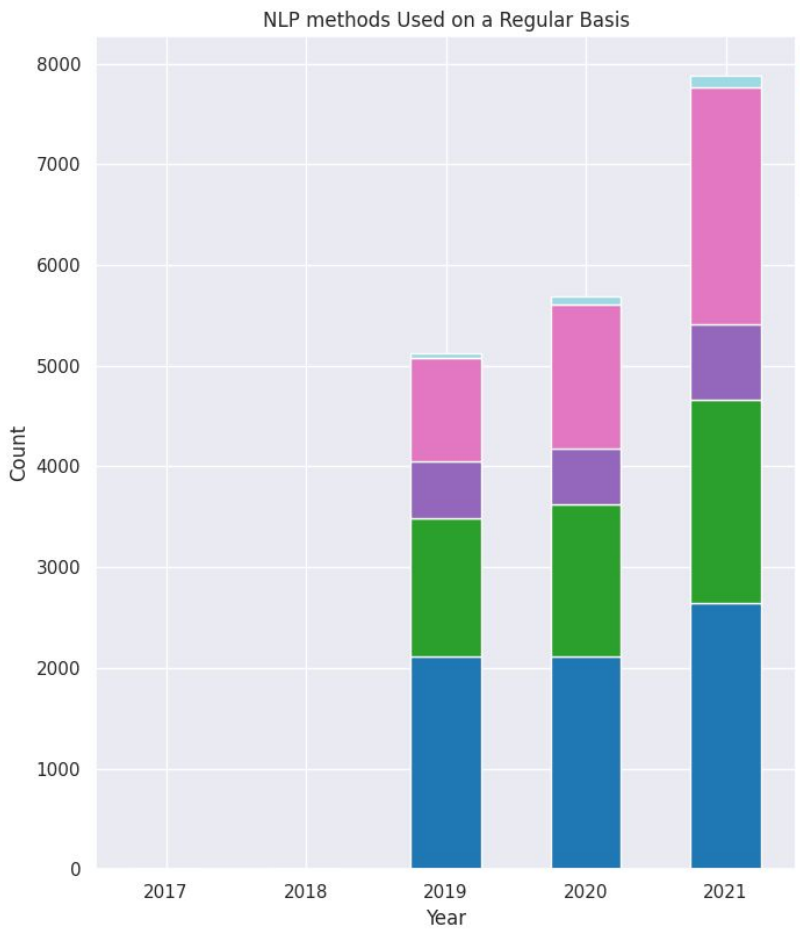


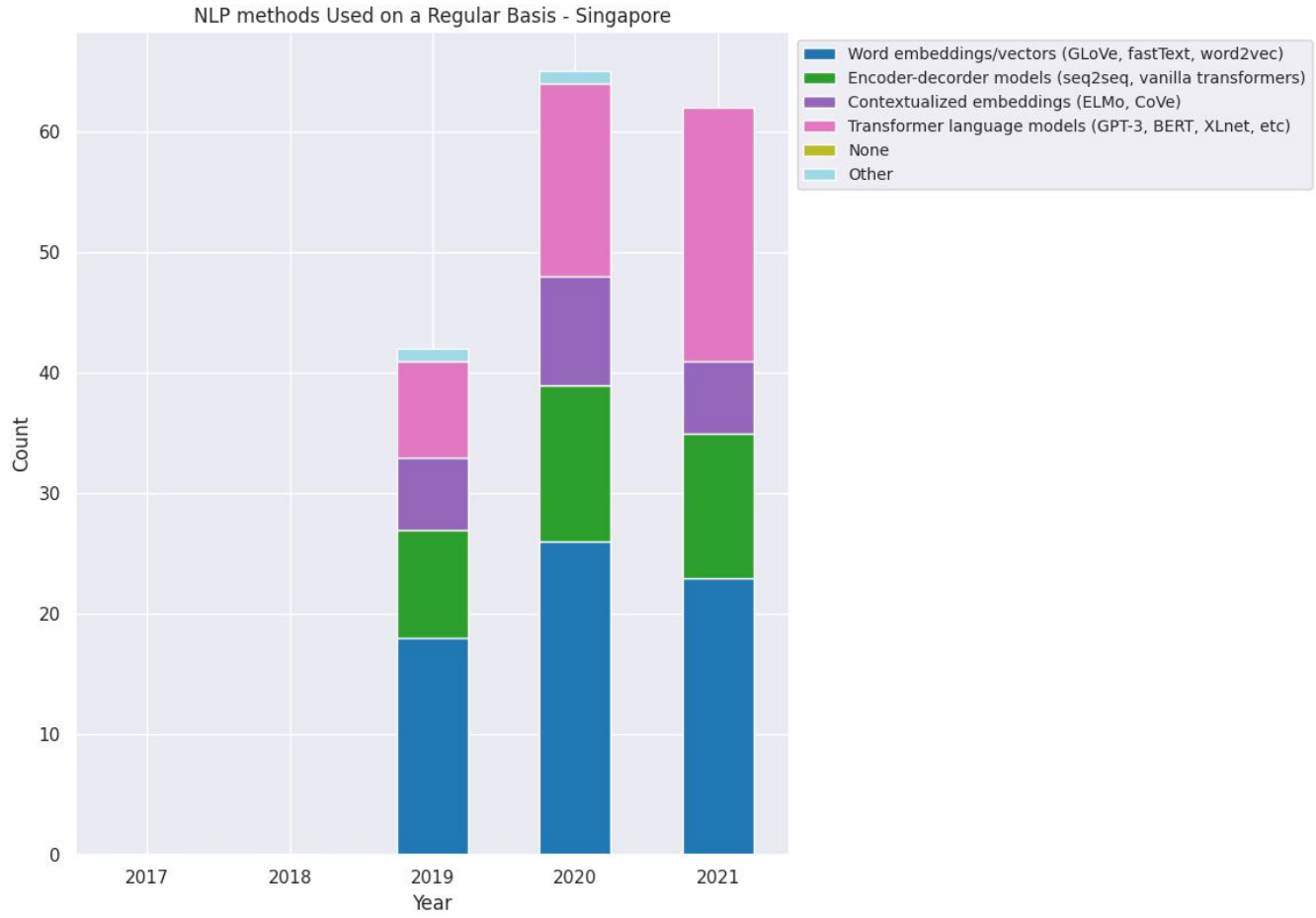


Insights:

- Most utilised Machine Learning algorithms include Regression Methods, Decision Trees, Boosting and Convolutional Neural Networks.
- In this case, the usage ML concepts are quite evenly spread out, which is unsurprising since different use cases require different solutions
- It might thus not be useful to focus too much on a single concept, but rather upskilling can take place at a higher level of critical thinking so that these ideas can be grasped quickly.

Domain Knowledge - NLP Methods

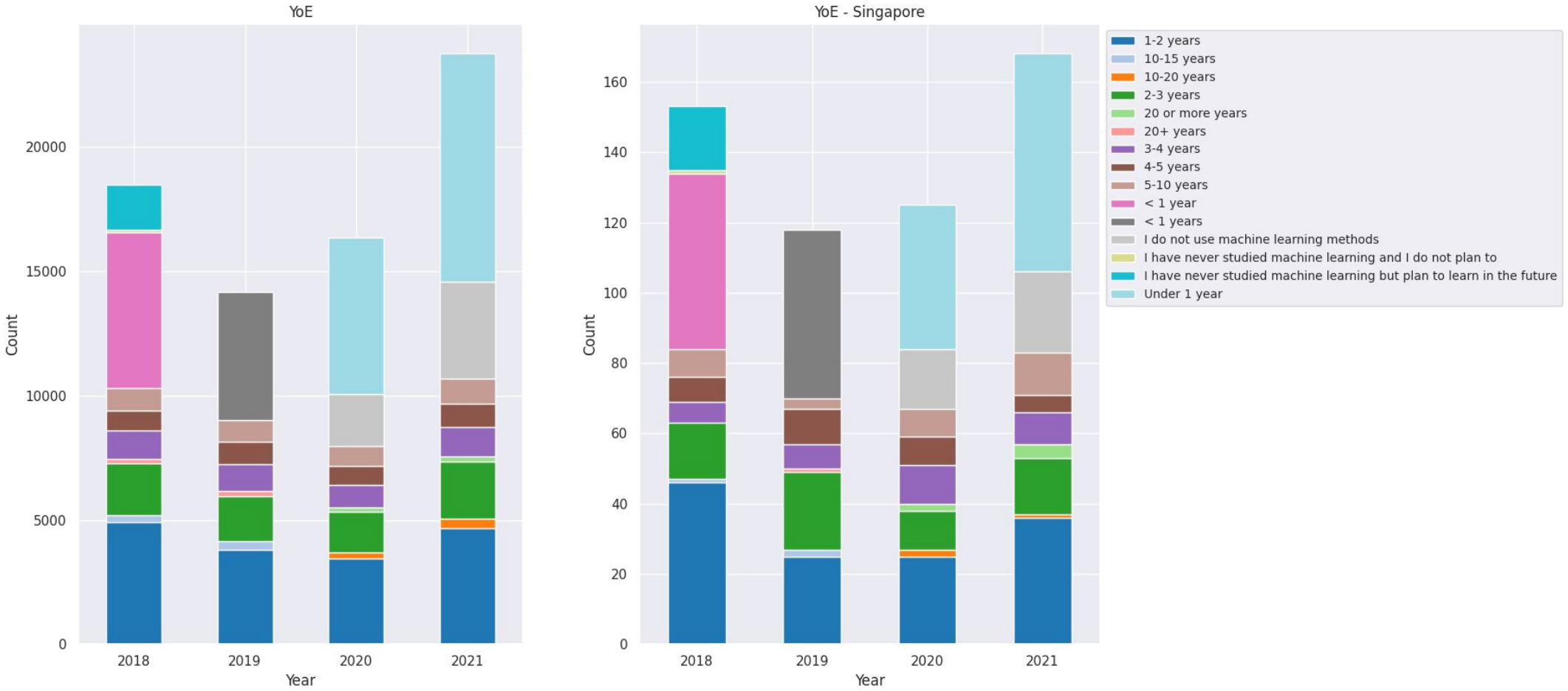




Insights:

- Word embeddings, encoder-decoder models and transformers remain popular
- Most notably, there has been a notable increase in usage of transformers
- SkillsFuture could potentially shift towards the area of transformers to cater to market demand

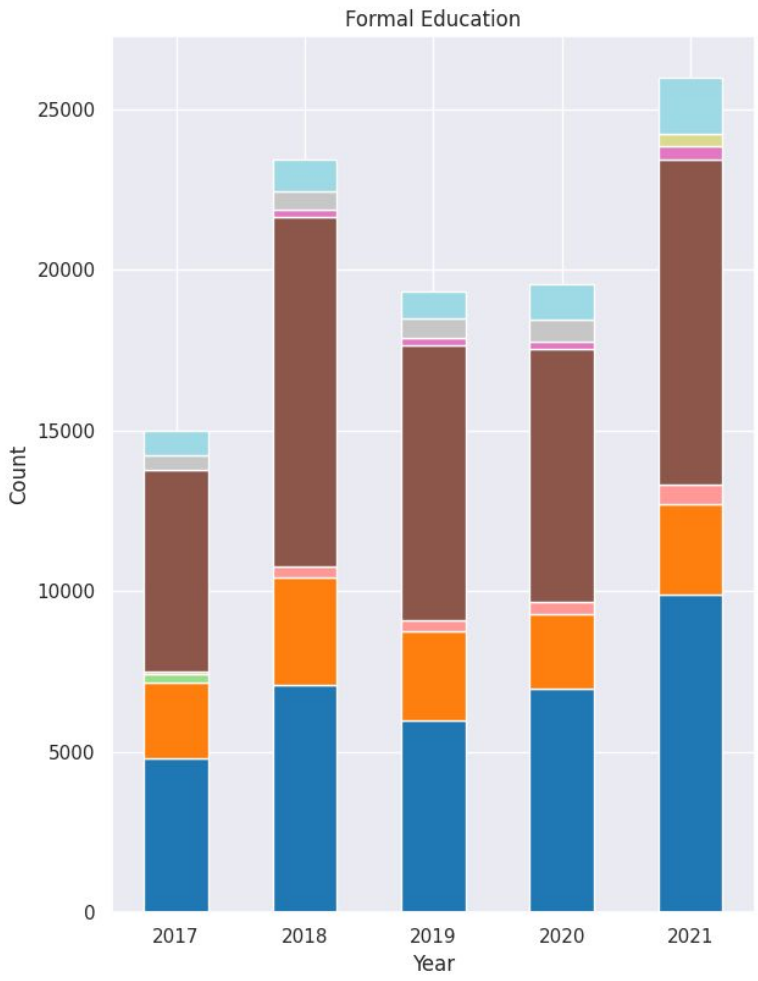
Domain Knowledge - Years of Experience

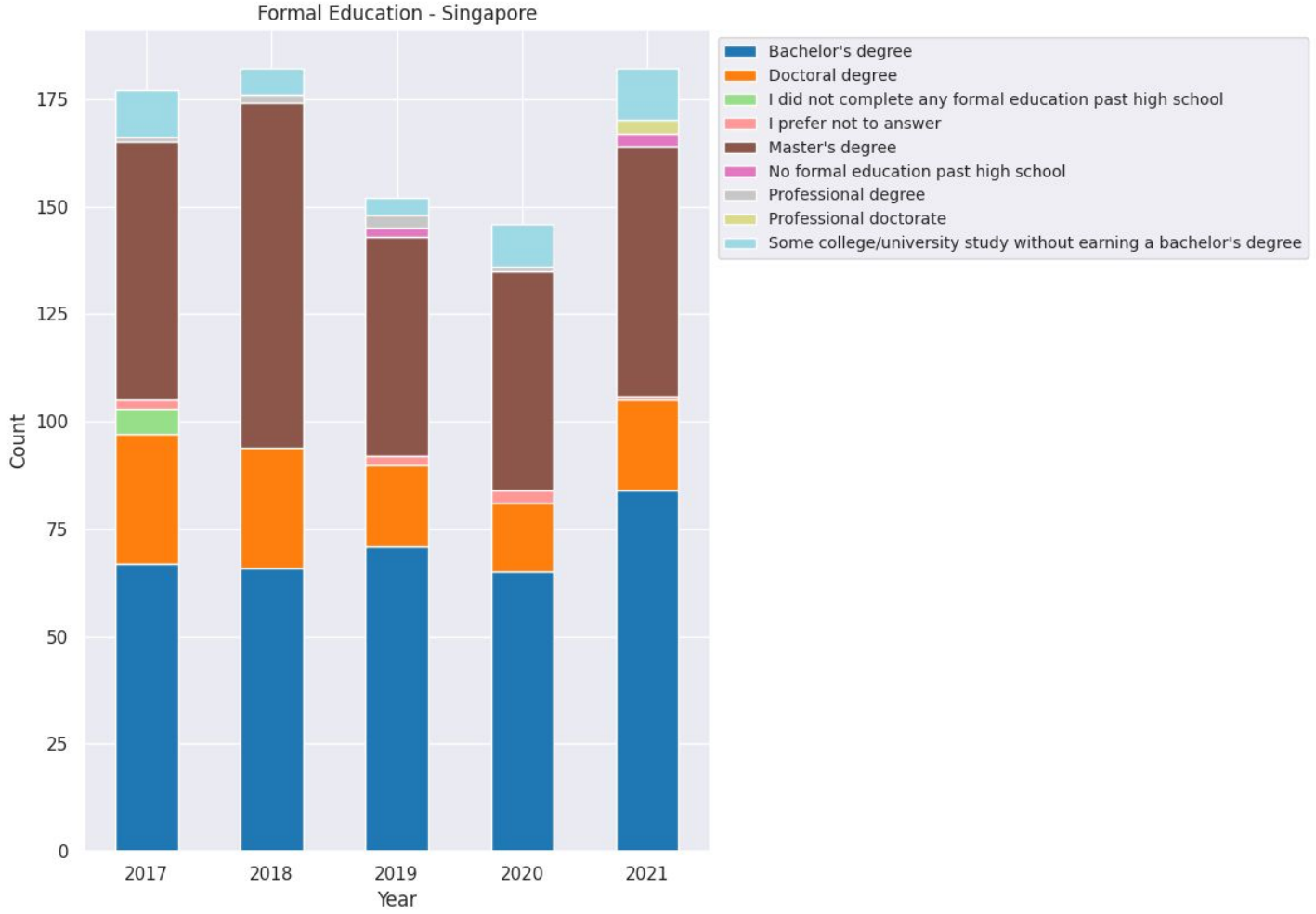


Insights:

- There is a clear increase in users that have under 1 year of experience, showing that there is a increase in the number of people who are interested in data analytic work
- SkillsFuture might have to increase their offerings in machine learning courses to meet demand as it seems like there will definitely be uptake in data analytic related courses

Domain Knowledge - Formal Education





Insights:

- Most users have a master's degree or a bachelor's degree
- Notably, the number of users with no formal education after high school has decreased to almost zero from 2017
- SkillsFuture could potentially offer courses to those who attended college but did not earn a bachelors or a bachelor's outside the data analytic field, since that is still a substantial portion.=

Section 2

Objective: To develop a fraud detection model to flag potentially fraudulent transactions for early intervention

Additionally, a model operational flow is to be implemented to monitor model performance and detect drifts

Background

Credit Card Fraud

- The Dataset has columns describing the credit card transaction, as well as a column describing if the case was an incident of fraud
- A test dataset and train dataset are provided

Additional Context:
Due to resource limitations, only 1000 manual investigations can be conducted monthly for ground truth establishment. This will determine the direction in which we build our model.

Model Building Approach

Monthly Statistics

Monthly statistics:

trans_date_trans_time	total_transactions	fraud_cases	fraud_rate_percent
2019-01	52525.0	506.0	0.963351
2019-02	49866.0	517.0	1.036779
2019-03	70939.0	494.0	0.696373
2019-04	68078.0	376.0	0.552308
2019-05	72532.0	408.0	0.562510
2019-06	86064.0	354.0	0.411322
2019-07	86596.0	331.0	0.382235
2019-08	87359.0	382.0	0.437276
2019-09	70652.0	418.0	0.591632
2019-10	68758.0	454.0	0.660287
2019-11	70421.0	388.0	0.550972
2019-12	141060.0	592.0	0.419680
2020-01	52202.0	343.0	0.657063
2020-02	47791.0	336.0	0.703061
2020-03	72850.0	444.0	0.609472
2020-04	66892.0	302.0	0.451474
2020-05	74343.0	527.0	0.708876
2020-06	57747.0	334.0	0.578385
Average	72037.5	417.0	0.609614

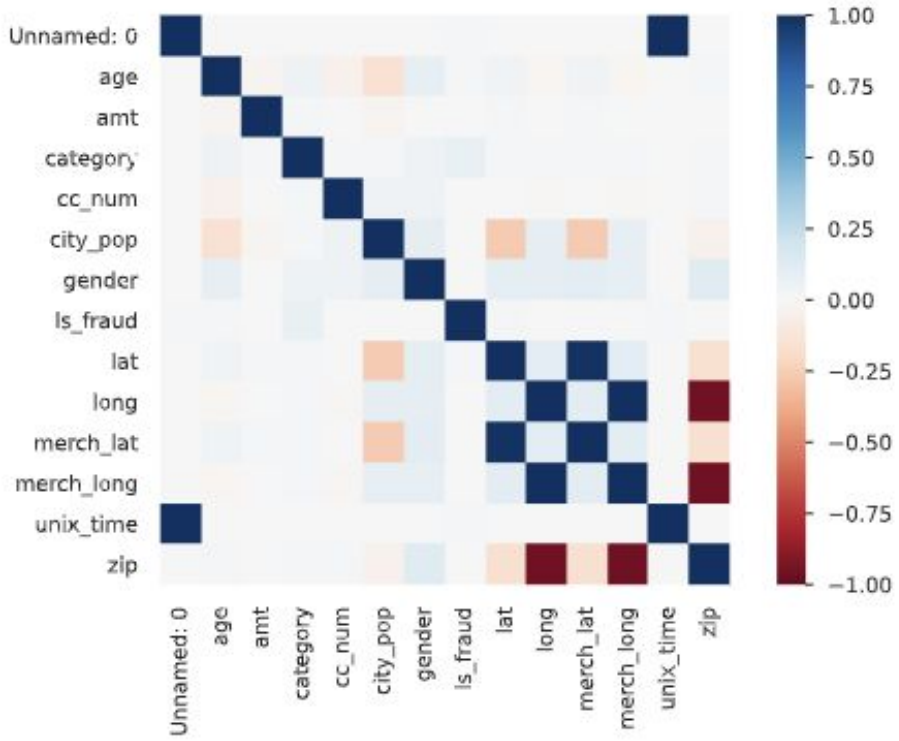
Overall fraud rate: 0.58%
Average monthly transactions: 72037.50
Average monthly fraud cases: 417.00
Average monthly fraud rate: 0.61%

If we prioritize recall (trying to catch all fraud cases):

We would generate too many alerts (including many false positives). With an investigation capacity of 1000 cases and 72,000+ transactions, We can only investigate about 1.4% of all transactions This would quickly exhaust your investigation resources on many non-fraudulent cases

Thus our model should focus on precision

EDA and Feature Selection



The correlation matrix shows that the following columns have high correlation: "long", "lat", "merch_long", "merch_lat", "zip", "unix_time","cc_num","city_pop" and are dropped

Column	Number of distinct values
trans_num	1,296,675
trans_date_trans_time	1,274,791

“trans_num” and “trans_date_trans_time” have extremely high cardinality and are dropped.

Additionally, date of birth, first name and last name are dropped as heuristically, they are not useful for fraud detection. Street is heuristically related to lat and long and is thus dropped as well.

Logistic Regression

Baseline Model

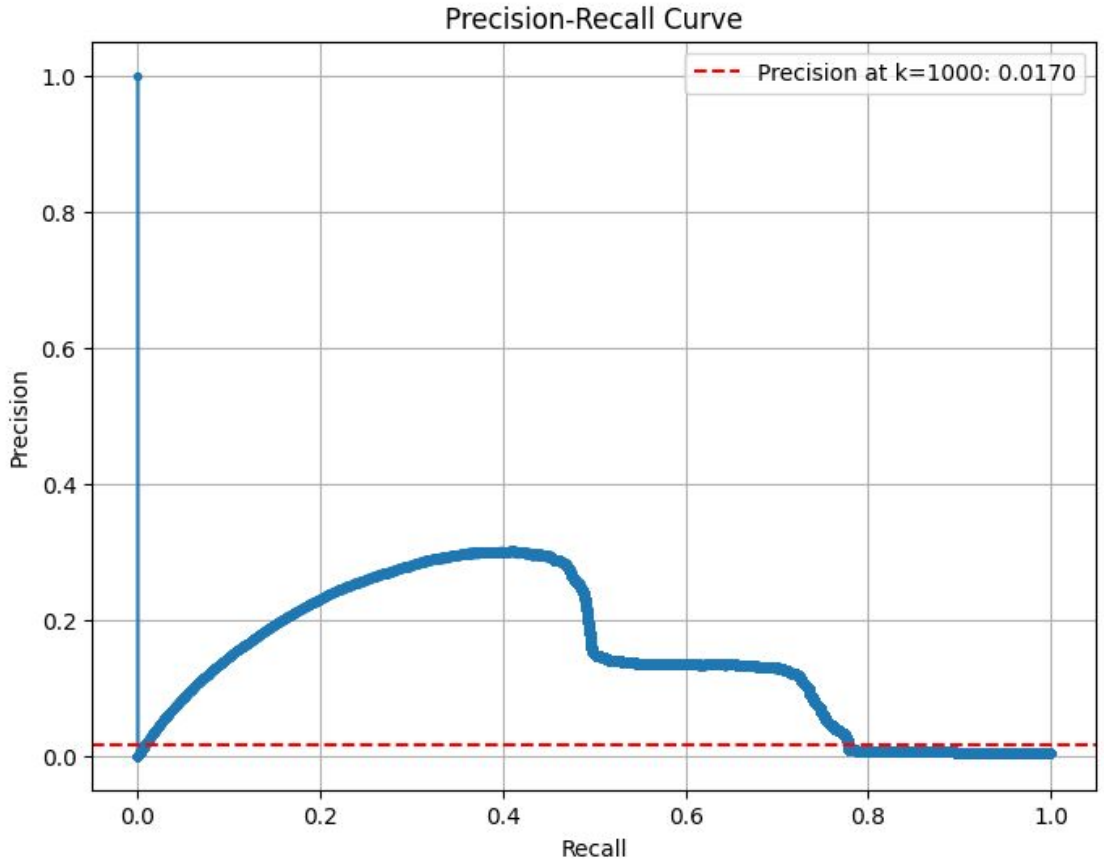
A threshold probability that would give approximately 1000 predictions per month is set. This threshold probability is then used to to investigate the top 1000 most suspicious transactions. The model metrics are then again assessed based on these 1000 cases.

Overall Model Results

	precision	recall	f1-score
0	1.00	0.95	0.97
1	0.06	0.75	0.10
accuracy			0.95
macro avg	0.53	0.85	0.54
weighted avg	1.00	0.95	0.97

```
Precision: 0.0555
Recall: 0.7524
```

Top 1000 Cases Results




```
Selected threshold: 0.9991
Number of alerts generated: 1000

Precision at k=1000: 0.0170
```

If we investigate your top 1000 most suspicious transactions, 17 would be actual fraud. This coupled with an overall low precision rules out this model

Neural Network

Model 1

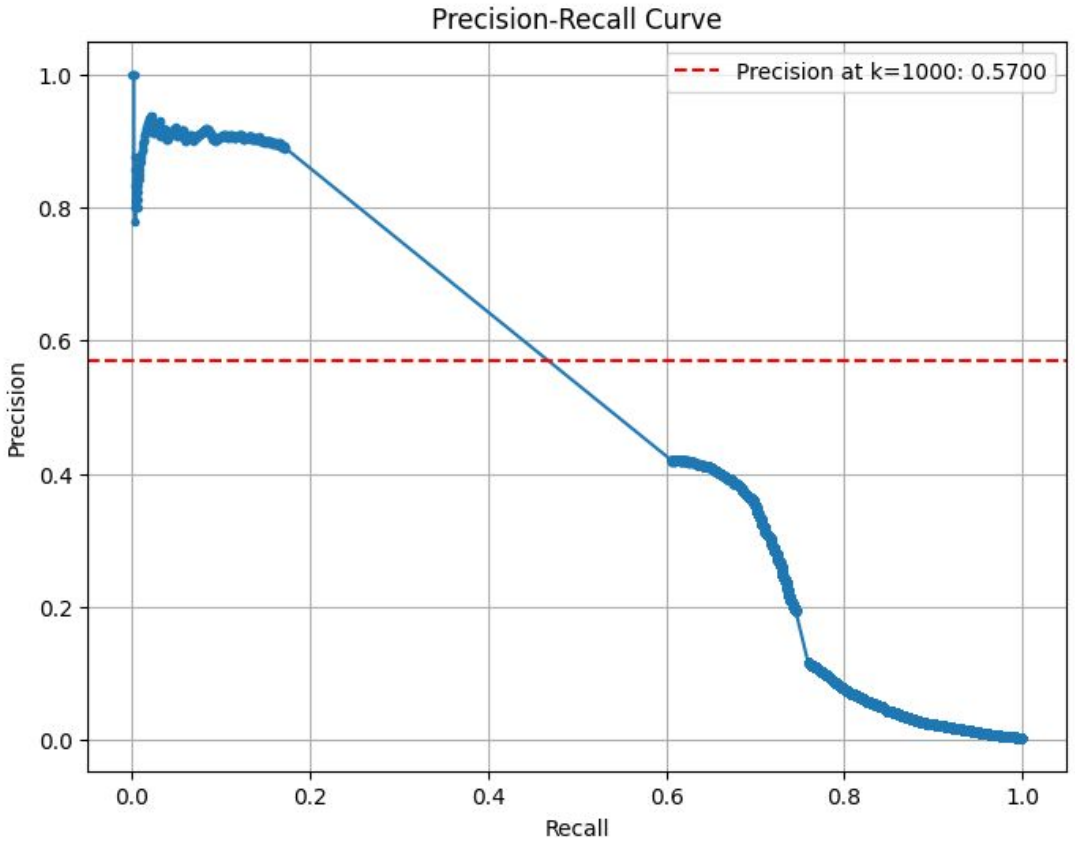
Neural networks utilise weights and biases in each node and each layer in order to adjust the weight of the function, eventually converging on an optimum weight for each node to produce the output. We aim to capitalise on this to analyze complex patterns to identify fraud

Overall Model Results

	precision	recall	f1-score
0	1.00	1.00	1.00
1	0.90	0.14	0.24
accuracy			1.00
macro avg	0.95	0.57	0.62
weighted avg	1.00	1.00	1.00

```
Precision: 0.9018
Recall: 0.1371
```

Top 1000 Cases Results




```
Selected threshold: 0.3915
Number of alerts generated: 3087

Precision at k=1000: 0.5700
```

Model performs significantly better. However, from the graph, we can see that performance drops off significantly as recall increases, indicating a lack of confidence about its bottom predictions. Additionally, recall overall is way too low

XGBoost

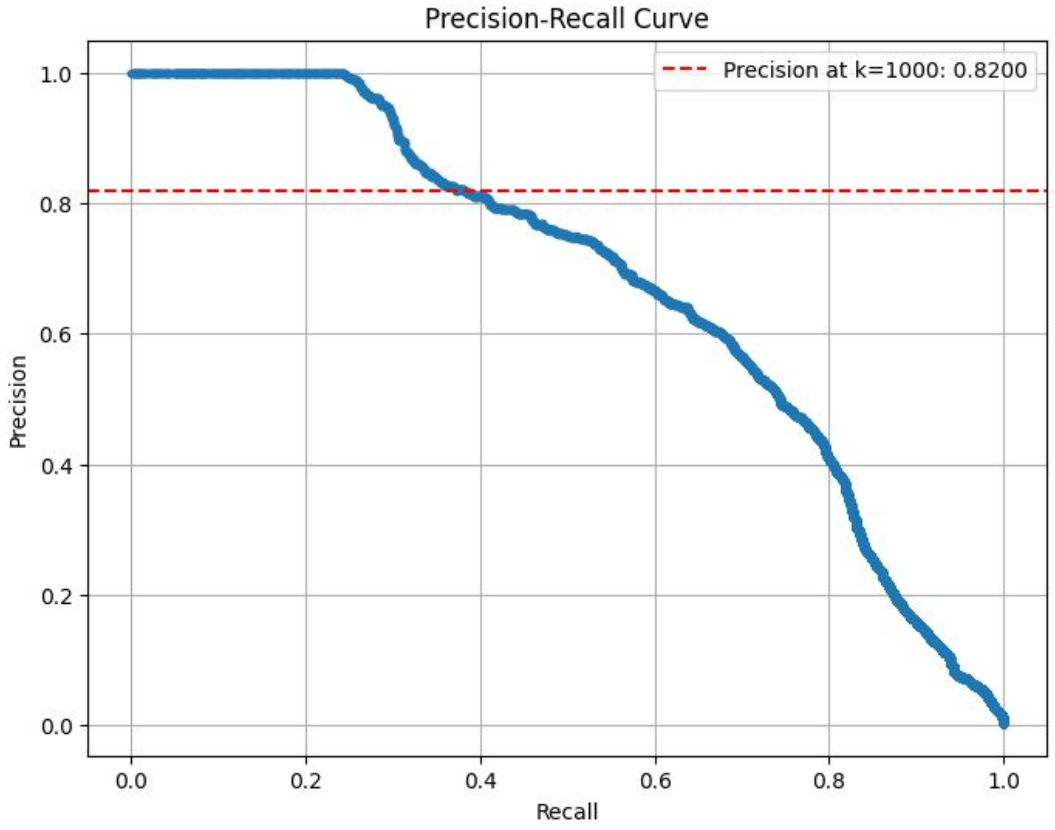
Model 2

Extreme gradient boosting is a form of boosting (ensemble learning), using decision trees as base learners to come to a 'democratic' decision. Each tree is able to correct for the mistakes of the previous tree in a sequential manner

Overall Model Results

	precision	recall	f1-score
0	1.00	1.00	1.00
1	0.72	0.55	0.62
accuracy			1.00
macro avg	0.86	0.77	0.81
weighted avg	1.00	1.00	1.00

Precision: 0.7239
Recall: 0.5464



Top 1000 Cases Results

Selected threshold: 0.6262
Number of alerts generated: 1000

Precision at k=1000: 0.8200

Final chosen model. For the 1000 most suspicious cases, 820 are actual fraud. The model is overall more confident even when recall is high, and is highly confident about it's top predictions, having a precision of 1.0 up to about 0.3 recall

Model Operational Flow

Model Deployment

Model Deployment:

- Deploy trained XGBoost model into production environment.
- Integrate the model into the transaction processing pipeline, where it evaluates transactions in real-time or in batches.
- The model will flag transactions as potentially fraudulent based on a prediction threshold (e.g., if probability > 0.5).

Model Prediction:

- Each transaction is passed through the model to receive a fraud prediction (fraud or not fraud) along with a confidence score (probability).
- Based on the model's output, flag the top transactions for manual investigation (up to 1000 per month as per the limit).

Feedback Loop:

- The manual investigation results (whether the flagged transaction was truly fraudulent or not) will be fed back into the system for model evaluation and retraining.
- This feedback will help to gradually improve the model's performance and ground truth labeling.

Model Operational Flow

Monitoring Model Performance and Metrics

Key Performance Metrics:

- Precision, Recall, and F1-Score: Measure the model's ability to correctly identify fraud while minimizing false positives (precision) and false negatives (recall). However, emphasis will still remain on precision.
- Confusion Matrix: Track the true positives, true negatives, false positives, and false negatives to understand how well the model is performing.
- Precision-Recall Curve: Similar to what was done previously, track the precision recall curve of the overall model and for the 1000 most suspicious cases.

Performance Monitoring:

- Monitor the model’s performance on an ongoing basis using metrics set out above
- Set thresholds for acceptable model performance and establish alerts for when metrics drop below acceptable levels (eg precision < 0.65)
- Regularly update the manual investigation process with new predictions and feedback to retrain the model periodically.

Monthly Evaluation:

- Each month, review the 1000 manually investigated transactions to ensure the model is still performing optimally.
- Track metrics over time and adjust thresholds if necessary to align with business goals.

Model Operational Flow

Model Drift Detection

Data Drift Detection:

- Feature distribution tracking: Monitor changes in input feature distributions over time (e.g., transaction amount, merchant, or customer demographics).
- Statistical tests (e.g., Kullback-Leibler divergence or Kolmogorov-Smirnov test) can be used to check for significant changes in feature distributions.
- Alert if input data deviates significantly from historical data patterns, indicating data drift that could impact model predictions.

Model Drift Detection:

- Track changes in model performance over time (e.g., precision, recall, or AUC) to detect model drift.
- Compare the model's predicted probabilities on recent data with those from the initial training dataset.
- Use concept drift detection methods, such as performance monitoring, to detect if the model's accuracy decreases over time.

Model Retraining:

- Set up an automated pipeline to trigger retraining when drift is detected, incorporating newly labeled data from manual investigations.
- Ensure the model is updated periodically (e.g., monthly or quarterly) to reflect changes in transaction patterns and fraud tactics.

Thank you