CreditByte

CreditByte, a mid-sized financial services company, has long been plagued by credit card fraud ever since AI systems have become more accessible. Their existing rule-based methods for detecting fraud – such as thresholding, unusual transaction amounts, or foreign transactions – seem to be outsmarted by modern-day criminals.

Because of this, the CEO of CreditByte, John Harlow, created an initiative for more intelligent solutions with their data. They have hired multiple different vendors, but they noticed that many of the solutions provided by the vendors were either not usable or misaligned to their business metrics. Apart from that, they found that it is hard to communicate with these vendors because many of what CreditByte wants is not as easily understood by these vendors.

So they decided to switch tactics: they decided to ask for proof-of-concept models first before committing to a full-fledged contract to different vendors. With this new strategy, they are able to find good quality vendors that they would be able to work harmoniously with. They prioritized having their data infrastructure ready and follows a medallion architecture to accommodate different analytics and business needs.

With these projects currently under development, they are finally looking into creating a model to address fraud. Slowly, John noticed that their data needs are becoming more complex and noticed that they lack the necessary data leadership within their own ranks. Because of this, he was recommended by one of his peers to hire Data Tinkers to create this model.

Data Tinkers

Data Tinkers is a startup data consultancy with a modest developer team size of 40, where most of the developers are data engineers and data architects. The company has an expertise in creating end-to-end data pipelines that provide solutions from data ingestion to predictive

analytics. While most of the developers are within the data infrastructure role, the company has a small data science team of 4 people, one of which is the lead named Mara.

They were recently hired by CreditByte to develop a proof-of-concept credit card fraud model. During the initial meeting, Mara noted that this may not have been the first time CreditByte hired vendors to create a model for them. She noticed that John Harlow sometimes slipped in mentioning about a previous effort that “didn’t push through”. In sensing this, Mara decided to create a proof-of-concept accurate enough but she wanted to prioritize on communicating with John and the rest of his team.

Mara was assigned this high-stakes project, and after some back-and-forth, noticed that their data on fraud cases was severely imbalanced. She got to work on the proof-of-concept while balancing her other projects and her wedding planning.

A twist in development

Midway through the development, CreditByte announced the hiring of a new Chief Data Officer, Jerome Corinthian. Jerome came from a cutthroat company and was looking at a new opportunity to put his experience to the test in the financial industry. Jerome, who’s eager to get things done, requested all current project statuses from all the vendors of CreditByte.

Jerome works fast and non-stop and he expects everyone that works with him to do the same. When he got hired, the number of meetings increased, deadlines were moved up, project get cancelled, and overall team dynamics was in disarray. He often stops presentations, not because he doesn’t understand the technical details, but because the presenter was taking too long to get to the point. Moreover, he if he is not convinced on the value of a project he would close the engagement abruptly as “it costs too much with little to no value for us.”

Mara was shocked to learn that their presentation was moved up to the same date as her wedding. To prepare for this, she asked you and the other data scientists to finish her current pipeline. Specifically, she wants you to update her current pipeline to handle the class imbalance in the dataset.

To help her out, you deprioritized your other projects and prepare for this one.

Mara’s strategy

She aligned you with her strategy for the proof-of-concept. Firstly, she was able to convince Jerome that the presentation would take place when she gets back. However, Jerome requested to have some reading materials in preparation for that meeting. Because of this, she tasked you in creating a report.

Secondly, she wants you to prioritize a high-level explanation to accommodate for Jerome. However, technical details should be provided. Jerome can understand all of the technical aspects of data science.

On the business metrics side, she noticed that CreditByte has been using two separate metrics that seem to be domain specific. These are, detection rate and fraud capture rate. She noticed that detection rate is the same as recall, while fraud capture rate follow the formula

$$\text{FCR} = \frac{\text{Total Number of fraudulent transactions flagged}}{\text{Total fraudulent transactions}} \times 100$$

She also wants you to provide how much money the model could save CreditByte to add more impact to the overall proof-of-concept. Because there's no set sampling strategy in place, she wants you to recommend the sampling strategy and justify the reason why in the report. She was wondering which would benefit the model best: class weighting, under-sampling, over-sampling, or a hybrid-approach. She ultimately wants to know what are the trade-offs from a technical and business standpoint.

Mara will ultimately read the report before she leaves, but don't expect that she would be as comprehensive.

Final thoughts

As you ponder on what Mara has told you and knowing a little bit on Jerome, you are placed in a position where you need to generate an output as close to what Mara would have created on her own.

You are ultimately tasked to produce a report that recommends a strategy on how to handle the class imbalance given the model. Within the report, it should justify the recommended strategy, show the typical model metrics, and highlights the potential business impact.

As you ponder further on the meeting with Mara, you are left to ask yourself "how would I create the report that would persuade Jerome, understandable to John, and is aligned with Mara's technical vision?"

Apart from that, what should be your main goal in writing this report? What is your main priority? What are some analysis that you could provide that could present the proof-of-concept in an impactful way? What are limitations that you can accept given the constraints? What is the overall scope of your work?

At the end of the meeting. You let out a sigh and think that maybe chaos like this is part of being a data scientist, and part of the job is to organize this chaos.