# 

# Industry Project Proposal

|  |  |
| --- | --- |
| Name of the Organization | Ibotta |
| Project location (city) | Denver, Colorado |

### 

### Project Description

**Problem definition**

*[50-100 word description of the problem which the candidates need to solve]*

|  |
| --- |
| *Ibotta’s* mission is to make all purchases rewarding via digital coupons. Unfortunately, fraudsters like to take advantage of our platform. We would like a model that:   1. Takes as input one more example accounts   and   1. Outputs one or more accounts that are very similar to the supplied accounts   This would allow our fraud investigators to quickly identify the same fraudulent activity in real-time and allow them to take action.  In this project the goal is: 1) To process customer account data and detect similarity to other accounts/transaction combinations in a scalable manner, 2) To process large volumes of data with minimal latency and high accuracy for fraud detection/outlier detection tasks. |

**Key Research Questions/ Technological constraints that the Project will Answer**

|  |
| --- |
| 1. Is it possible to detect specific text/patterns from transaction data streams? 2. Is it possible to uniquely identify outliers of fraudulent transactions and accounts with minimal latency? 3. Is it possible to cluster user accounts based on “similarity” in transaction types for efficient account monitoring and promotional offers. |

**Final deliverables at the end of the project**

*[Please list the desired technical deliverables from the project team in as much detail as possible]*

|  |
| --- |
| 1. End-to-end pipeline that identifies fraudulent accounts. 2. Quantitative and qualitative ways to identify fraudulent accounts and similar accounts. 3. Deployed ML model 4. Technical report on System setup and performances. |

**Key activities/ technologies the project team may be expected to undertake/ work with**

*[E.g. What kind of technology stack they will work with, the datasets they may need to work on, what kind of analysis they may be expected to undertake, etc.]*

|  |
| --- |
| 1. Process Dependencies:   Python  Tensorflow or pytorch  Flask or FastAPI front end  Deployment on AWS or Google cloud.  Optical character recognition system   1. Public data (You may choose any 2):   i. Financial Data set <https://web.archive.org/web/20161019192412/http://lisp.vse.cz/pkdd99/berka.htm> (\*note: likely not available, code is there in GitHub but data link broken. However, site provides useful way to look at the problem)  Accompanying code: <https://github.com/justinng1/berka>  ii. Bank Transaction Data  <https://www.kaggle.com/apoorvwatsky/bank-transaction-data>  iii. EU Financial Transactions Dataset:  <https://data.europa.eu/euodp/data/dataset/zhr959EjfKSPU6UBmQaITg/resource/ee6050a1-e6ea-429c-9199-994b42505b24>    <https://www.kaggle.com/mkechinov/ecommerce-purchase-history-from-electronics-store> |

**Expected learning outcomes**

*[What do you expect the candidates to learn from the project. Please mention the technical skills they will imbibe over the project.]*

|  |
| --- |
| 1. Building end-to-end ML pipelines. 2. Qualitative and quantitative reporting for fraudulent and similar account detection. 3. ML Model development to train models on how to detect fraudulent and similar accounts. 4. Communication and technical reporting. |