**6242- Data Visualization (Project)**

**Data link:** [**https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?resource=download**](https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?resource=download)

**Proposal & Presentation - 10/13 [12.5%]**

1. Latex Documentation
2. 9 Heilmeier questions ([source](http://en.wikipedia.org/wiki/George_H._Heilmeier))
   1. What are you trying to do? Articulate your objectives using absolutely no jargon.
   2. How is it done today; what are the limits of current practice?
   3. What's new in your approach? Why will it be successful?
   4. Who cares?
   5. If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
   6. What are the risks and payoffs?
   7. How much will it cost?
   8. How long will it take?
   9. What are the midterm and final "exams" to check for success? How will progress be measured?

**Progress Report - 11/3 [5%]**

**Final Poster presentation & Final report - 12/1 [32.5%]**

1. Back- end Code (Python) - Data Cleaning, Visualization, Modelling, PCA, Feature Engineering & Selection, Implementation from Research Paper, UI - Streamlit
2. User Interface(Streamlit)
   1. Modeling/Clustering/Analysis - Viz
   2. Predictive Analytics for Detecting Fraud
3. Data Visualization - Plotly
4. Data Backend RDBMS - SQLITE
5. Versioning - Github GeorgiaTech.
6. Presentation - PowerPoint
7. Data : /kaggle/input/bank-account-fraud-dataset-neurips-2022/Base.csv

**Task**

1. Data Analysis - Report
2. Modeling , Feature Importance ,Selection
3. API Build
4. Build UI
5. Build Backend
6. Integration

—------------------------------------

Piyush writing this

* 1. What are you trying to do? Articulate your objectives using absolutely no jargon.

We are creating a web based tool to prevent banking fraud. We aim to find the insights from transactional banking data, and using unsupervised machine learning algorithms such as k-means clustering develop this tool which would be able to prevent the fraudulent transaction by generating alerts and in case if the fraud was not caught in time, then detect it faster than traditional methods so that scale of the fraud can be controlled.

* 1. How is it done today; what are the limits of current practice?
  2. What's new in your approach? Why will it be successful?
  3. Who cares?
  4. If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
  5. What are the risks and payoffs?
  6. How much will it cost?
  7. How long will it take?
  8. What are the midterm and final "exams" to check for success? How will progress be measured?

—------------------------------------------------

**# Updated by Jagan (9/11/2023)**

**Project Proposal: Detecting Fraudulent Transactions in Bank Accounts**

**1. What are you trying to do?**

We aim to develop a web application that can identify and flag fraudulent transactions in bank accounts. Our goal is to create a user-friendly and efficient system that helps financial institutions and individuals quickly detect and respond to suspicious activities without requiring extensive technical knowledge.

**2. How is it done today, and what are the limits of current practice?**

Currently, fraud detection in banking relies heavily on rule-based systems and manual review by fraud analysts. These methods have several limitations:

- Limited scalability: Manual reviews are time-consuming and cannot handle the increasing volume of transactions.

- Reactive approach: Current systems often detect fraud after the fact, leading to financial losses.

- False positives: Rule-based systems can generate many false alarms, causing inconvenience to customers.

- Inefficient use of resources: Fraud analysts spend significant time on routine checks, which could be automated.

**3. What's new in your approach and why do you think it will be successful?**

Our approach leverages advanced machine learning algorithms to automate the detection of fraudulent transactions. Key innovations include:

- Machine Learning Models: We will use state-of-the-art machine learning techniques to create models that can adapt and learn from new fraud patterns.

- Real-time Processing: Our system will process transactions in real-time, allowing for immediate response to suspicious activities.

- Reduced False Positives: By analyzing transaction patterns and customer behavior, we aim to reduce false positive alerts.

- User-Friendly Interface: The web application will be designed for easy integration into existing systems and will provide user-friendly dashboards for monitoring and decision-making.

**4. Who cares? If you're successful, what difference will it make?**

Our project will benefit various stakeholders:

- Financial Institutions: Banks and credit card companies can reduce fraud-related losses and enhance customer trust.

- Individuals: Customers will have better protection against unauthorized transactions.

- Regulators: Compliance with regulations related to fraud detection and prevention will be improved.

- Society: Reducing financial fraud contributes to a safer and more secure financial ecosystem.

**5. What are the risks and the payoffs?**

Risks:

- Data Privacy: Ensuring the security and privacy of customer data is critical.

- Model Accuracy: The success of the project depends on the accuracy of the machine learning models.

Payoffs:

- Reduced Fraud Losses: Financial institutions can save millions by preventing fraud.

- Enhanced Customer Trust: Customers will have more confidence in their banks.

- Competitive Advantage: Banks implementing our system can gain a competitive edge.

- Improved Compliance: Meeting reguTlatory requirements will avoid penalties.

**6. How much will it cost?**

The cost of the project will depend on various factors, including the complexity of the system, data volume, and the size of the development team. We will provide a detailed cost estimate as part of the project plan.

**7. How long will it take?**

The project timeline will be finalized after a detailed analysis, but we estimate that it will take approximately 12-18 months for development, testing, and deployment.

8. What are the midterm and final "exams" to check for success?

Midterm Exam:

- Successfully develop and train machine learning models on a representative dataset.

- Achieve a reduction in false positives and demonstrate improved fraud detection.

Final Exam:

- Deploy the web application in a real banking environment.

- Monitor the system's performance and assess its effectiveness in reducing fraud.

- Gather user feedback and make necessary adjustments for optimization.

**Project Plan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Start Date** | **End Date** | **Assigned To** | **Comment** |
| Proposal | 18-Sep | 29-Sep | Nukta & Emir | Original Submission Date - 10/13 |
| Data Analysis & Visualization | 30-Sep | 7-Oct | Nukta & Piyush |  |
| Data Pre-Processing & Modelling | 8-Oct | 28-Oct | Jagannath & Ashish |  |
| Progress Report | 29-Oct | 3-Nov | Emir & Piyush |  |
| Building Backend & Front End | 4-Nov | 11-Nov | Jagannath & Emir |  |
| Testing | 12-Nov | 18-Nov | Piyush & Ashish |  |
| Integration | 19-Nov | 25-Nov | Emir & Piyush |  |
| Final Report & Poster Presentation | 26-Nov | 28-Nov | Jagannath/Nukta/Ashish/Emir/Piyush | Original Submission Date - 12/1 |

**Literatures**

|  |  |  |  |
| --- | --- | --- | --- |
| **S\_No** | **Paper** | **Author** | **Link** |
| **1** | **Credit Card Fraud Detection Using Machine Learning: A Survey** | **Yvan Lucas, Johannes Jurgovsky** | [**https://arxiv.org/abs/2010.06479**](https://arxiv.org/abs/2010.06479) |
| **2** | **A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective** | **SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi** | **https://arxiv.org/abs/1611.06439** |
| **3** | **Deep Learning for Anomaly Detection: A Survey** | **Raghavendra Chalapathy;Sanjay Chawla** | **https://arxiv.org/abs/1901.03407** |
| **4** | **Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances** | **Waleed Hilal, S. Andrew Gadsden, John Yawney** | **https://www.sciencedirect.com/science/article/pii/S0957417421017164** |
| **5** | **Survey of fraud detection techniques** | **Yufeng Kou; Chang-Tien Lu; S. Sirwongwattana; Yo-Ping Huang** | **https://ieeexplore.ieee.org/document/1297040** |
| **6** | **Survey on Anomaly Detection using Data Mining Techniques** | **Shikha Agrawal, Jitendra Agrawal** | [**https://www.sciencedirect.com/science/article/pii/S1877050915023479**](https://www.sciencedirect.com/science/article/pii/S1877050915023479) |
| **7** | **Robust Random Cut Forest Based Anomaly Detection On Streams** | **Sudipto Guha;Nina Mishra;Gourav Roy;Okke Schrijvers** | **https://proceedings.mlr.press/v48/guha16.pdf** |
| **8** | **A Survey of Outlier Detection Methodologies** | **Victoria J. Hodge;Jim Austin** | [**https://www-users.york.ac.uk/~vjh5/myPapers/Hodge+Austin\_OutlierDetection\_AIRE381.pdf**](https://www-users.york.ac.uk/~vjh5/myPapers/Hodge+Austin_OutlierDetection_AIRE381.pdf) |
| **9** | **A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach** | **Toluwase Ayobami Olowookere, Olumide Sunday Adewale** | [**https://www.sciencedirect.com/science/article/pii/S2468227620302027**](https://www.sciencedirect.com/science/article/pii/S2468227620302027) |
| **10** | **Intelligent financial fraud detection: A comprehensive review** | **Jarrod West, Maumita Bhattacharya** | **https://www.sciencedirect.com/science/article/abs/pii/S0167404815001261** |
| **11** | **How to detect healthcare fraud? “A systematic review”** | **Andi Yaumil Bay R. Thaifur , M. Alimin Maidin , Andi Indahwaty Sidin , Amran Razak** | [**https://www.sciencedirect.com/science/article/pii/S0213911121002661?via%3Dihub**](https://www.sciencedirect.com/science/article/pii/S0213911121002661?via%3Dihub) |
| **12** | **Credit Card Fraud Detection using Machine Learning: A Systematic Literature Review** | **Harish Paruchuri** | [**https://pdfs.semanticscholar.org/7bed/e29abeb0facd7a1c1754be2e6e1d3668ab9c.pdf**](https://pdfs.semanticscholar.org/7bed/e29abeb0facd7a1c1754be2e6e1d3668ab9c.pdf) |
| **13** | **Credit Card Fraud Detection: A Systematic Review** | **Victoria Priscilla & Padma Prabha** | [**https://link.springer.com/chapter/10.1007/978-3-030-38501-9\_29**](https://link.springer.com/chapter/10.1007/978-3-030-38501-9_29) |
| **14** | **A Comprehensive Survey of Data Mining-based Fraud Detection Research** | **CLIFTON PHUA, VINCENT LEE, KATE SMITH, & ROSS GAYLER** | **https://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf** |
| **15** | **Review of Machine Learning Approach on Credit Card Fraud Detection** | **Rejwan Bin Sulaiman, Vitaly Schetinin & Paul Sant** | [**https://link.springer.com/article/10.1007/s44230-022-00004-0**](https://link.springer.com/article/10.1007/s44230-022-00004-0) |

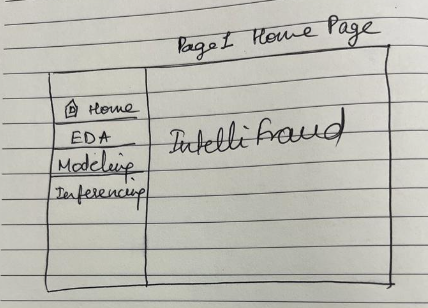
|  |  |
| --- | --- |
| Paper | Approach |
| Credit Card Fraud Detection Using Machine Learning: A Survey | Survey of different approaches random forest, neural networks, SVM,RNN,LSTM  logistic regression classifier, Sequence Modelling, Hidden Markov Chain |
| A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective | Survey - ANN, GNN, Hidden Markov Chain, SVM, Bayesian, Fuzzy Logic, Fuzzy NN,Genetic Algorithm, Decision Tree,Case Based Reasoning |
| Deep Learning for Anomaly Detection: A Survey | Deep Learning/ Deep Anomaly Detection (CNN+RNN+LSTM) |
| Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances | SVM, CNN+RNN+LSTM |
| Survey of fraud detection techniques | GNN, Rule Based, Visualizations |
| Survey on Anomaly Detection using Data Mining Techniques | K-Means, K-Medoids, EM Clustering, Outlier detection Algorithm, Tree, Fuzzy Logic, Naïve Bayes, Genetic Algorithm, NN, SVM, Hybrid Approach (DT & SVM) |
| Robust Random Cut Forest Based Anomaly Detection On Streams | Isolation Forest |
| A Survey of Outlier Detection Methodologies | Proximity Based Techniques (KNN); Convex Peeling, PCA, KNN+K-Mean, ANN |
| A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach | Cost Sensitive Meta Lavel Learning Algorithm (K-Nearest Neighbour, Decision Tree, and Multilayer Perceptron algorithms) |
| Intelligent financial fraud detection: A comprehensive review |  |
| How to detect healthcare fraud? “A systematic review” |  |
| Credit Card Fraud Detection using Machine Learning: A Systematic Literature Review |  |
| Credit Card Fraud Detection: A Systematic Review |  |
| A Comprehensive Survey of Data Mining-based Fraud Detection Research |  |
| Review of Machine Learning Approach on Credit Card Fraud Detection  Data Analysis and Visualization  Modeling  Building Back end  Building Front  Literature Survey |  |

**Project Execution**

Overall User Interface - 3 Pages (Use Streamlit Pages) : <https://docs.streamlit.io/library/get-started/multipage-apps/create-a-multipage-app>

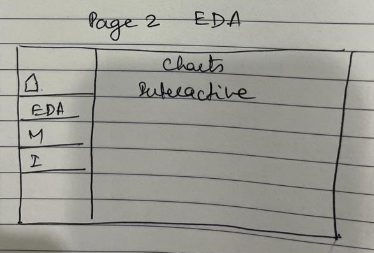
* Home
* EDA
* Modelling
* Inferencing

Home Page

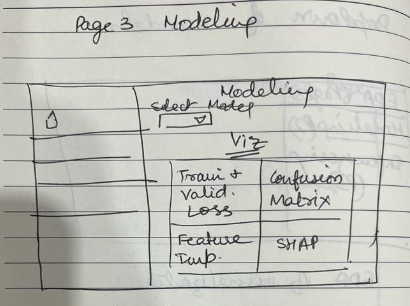


**EDA Page**

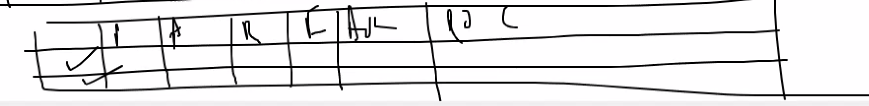
**Details…..**

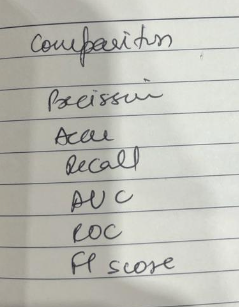
****

**Modelling Page**

****

The below are model metrics that will be added below graphs

****

****

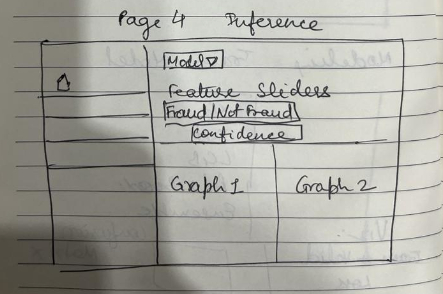
**Input : Dropdown and Submit Button**

|  |
| --- |
| **Model (Drop Down)** |
| **XGBoost** |
| **LGB** |
| **AdaBoost** |
| **CatBoost** |
| **Ensemble** |

**Model Metrics (Viz)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Train/Validation Loss** | **Confusion Matrix** | **Feature Importance** | **SHAP** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Inferencing Page**

****

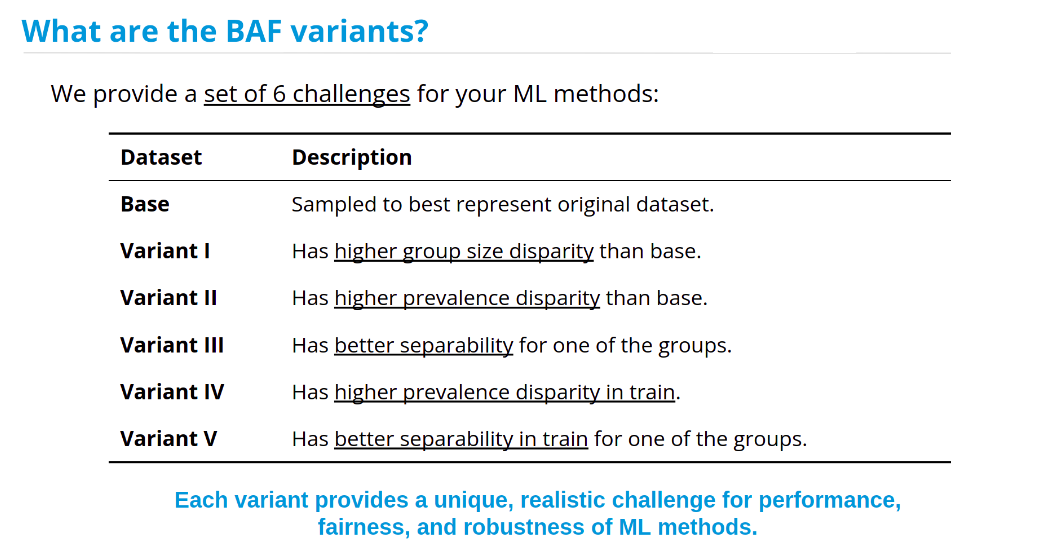
Model - Dropdown

Top Features as (5 to 7) as slider/filter as needed

Submit Button

Inferencing Visualization (Research) - Fraud/Not Fraud ; Confidence (Predict\_Proba) ; Others …research

**Data Variants**

****

**A graph of fraud and not fraud

Description automatically generated**

**Features Selection:**

|  |  |  |
| --- | --- | --- |
| **Action** | **Method** | **Features Removed** |
| **Remove low variance feature** | **Variance Threshold** | **device\_fraud\_count** |
| **Remove highly correlated features** | **Pearson's Correlation Matrix** | **velocity\_4w** |
| **Extract from model** | **Feature Importance’s** |  |
|  |  |  |
|  |  |  |

**Final Feature Selected:**

housing\_status

device\_os

credit\_risk\_score

current\_address\_months\_count

has\_other\_cards

keep\_alive\_session

prev\_address\_months\_count

phone\_home\_valid

proposed\_credit\_limit

name\_email\_similarity

income

**A grid with text and numbers

Description automatically generated with medium confidence**

**A graph with blue and white text

Description automatically generated**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample\_Size | Classifier | Conf\_Mtrx | Accuracy | Precision | Recall | F1\_Score | ROC\_AUC\_Scr | CV\_Score |
| 1:01 | XGBClassifier | [[10280 2855]  [ 2865 10471]] | 0.784 | 0.786 | 0.785 | 0.785 | 0.784 | 0.783 |
| 1:01 | AdaBoostClassifier | [[10411 2724]  [ 2981 10355]] | 0.784 | 0.792 | 0.776 | 0.784 | 0.785 | 0.785 |
| 1:01 | LGBMClassifier | [[10229 2906]  [ 2894 10442]] | 0.781 | 0.782 | 0.783 | 0.783 | 0.781 | 0.782 |
| 1:01 | VotingClassifier | [[10337 2798]  [ 2887 10449]] | 0.785 | 0.789 | 0.784 | 0.786 | 0.785 | 0.787 |
| 1:01 | StackingClassifier | [[10336 2799]  [ 2882 10454]] | 0.785 | 0.789 | 0.784 | 0.786 | 0.785 | 0.787 |
| 1:02 | XGBClassifier | [[23294 3281]  [ 4604 8528]] | 0.801 | 0.722 | 0.649 | 0.684 | 0.763 | 0.797 |
| 1:02 | AdaBoostClassifier | [[23773 2802]  [ 4968 8164]] | 0.804 | 0.744 | 0.622 | 0.678 | 0.758 | 0.797 |
| 1:02 | LGBMClassifier | [[23755 2820]  [ 5083 8049]] | 0.801 | 0.741 | 0.613 | 0.671 | 0.753 | 0.794 |
| 1:02 | VotingClassifier | [[23538 3037]  [ 4778 8354]] | 0.803 | 0.733 | 0.636 | 0.681 | 0.761 | 0.799 |
| 1:02 | StackingClassifier | [[23516 3059]  [ 4771 8361]] | 0.803 | 0.732 | 0.637 | 0.681 | 0.761 | 0.799 |
| 1:03 | XGBClassifier | [[36534 3276]  [ 6012 7120]] | 0.825 | 0.685 | 0.542 | 0.605 | 0.730 | 0.820 |
| 1:03 | AdaBoostClassifier | [[37160 2650]  [ 6372 6760]] | 0.830 | 0.718 | 0.515 | 0.600 | 0.724 | 0.819 |
| 1:03 | LGBMClassifier | [[37323 2487]  [ 6625 6507]] | 0.828 | 0.723 | 0.496 | 0.588 | 0.717 | 0.816 |
| 1:03 | VotingClassifier | [[36984 2826]  [ 6216 6916]] | 0.829 | 0.710 | 0.527 | 0.605 | 0.728 | 0.821 |
| 1:03 | StackingClassifier | [[36929 2881]  [ 6145 6987]] | 0.830 | 0.708 | 0.532 | 0.608 | 0.730 | 0.821 |

**Performance ROC-AUC & Precision Recall Curve at 1:1 Ratio**

**A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated**

**A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated**

**A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated**

**A diagram of a network

Description automatically generated**