IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Fraud Detection System Employing Machine Learning Techniques for Credit Card Transactions

G V S Abhishek Varma[1], G V N Akshay Varma[2], Adari Kethaan[3]
Student CSE Department, Maharaj Vijayaram Gajapathi Raj College of Engineering, AP, India[1,3]
Student CSE Department, Sri Sivasubramaniya Nadar College of Engineering, TN, India[2]

Abstract: *This research paper presents a comprehensive demonstration of the Microsoft Fabric data science workflow, illustrating its effectiveness through an end-to-end example centered on the development of an advanced fraud detection system. The primary objective of this study is to construct a robust and efficient fraud detection mechanism by leveraging the capabilities of machine learning algorithms trained on historical data encompassing instances of fraudulent activities. The overarching aim is to discern intricate patterns inherent in fraudulent events, thus empowering the system to swiftly and accurately identify and flag such activities in case of their recurrence. This paper expounds upon a meticulously designed workflow that encompasses a series of pivotal steps, including the installation of custom libraries tailored to the task, meticulous data loading and preprocessing, a comprehensive exploratory data analysis phase aimed at extracting meaningful insights, the intricate process of training a machine learning model using Scikit-Learn and Flow, the critical step of selecting and registering the most performant model, and, finally, the seamless deployment of the trained model for real-time scoring and prediction.The core of the presented workflow centers on the concept of a lake house, wherein the data is sourced from a public blob and subsequently stored for comprehensive analysis. This architectural paradigm underscores the significance of unified data storage, offering a coherent platform for seamless integration and manipulation. The research paper emphasizes that this approach not only elevates the efficiency of data handling but also lays the foundation for consistent and structured analysis, ultimately enhancing the accuracy and applicability of the subsequent machine learning stages.By tackling the complexities of a real-world scenario involving fraud detection, this research paper underscores the versatility and adaptability of the Microsoft Fabric framework. While the primary focus is on the development of a robust fraud detection system, the significance of this approach reverberates across diverse data science endeavors. The intricate process of custom library installation, data ingestion, preprocessing, and model deployment provides an invaluable resource for practitioners navigating the multifaceted landscape of data science workflows within the Microsoft Fabric ecosystem.In a world characterized by an exponential surge in data generation and a corresponding demand for actionable insights, the presented paper serves as more than just a solution to the specific problem of fraud detection. It morphs into a foundational template, empowering data scientists, researchers, and industry professionals with a structured approach to harnessing the potential of complex datasets. The step-by-step elucidation of the workflow equips practitioners with a tangible guide, offering insights into handling intricacies that often accompany large-scale data analysis. Through its detailed exposition, the paper bridges the gap between theory and practice, allowing readers to not only comprehend the theoretical underpinnings of the Microsoft Fabric framework but also practically implement and adapt its methodologies to suit a myriad of data science challenges. In conclusion, this research paper transcends the confines of a traditional study on fraud detection by encompassing a holistic exploration of the Microsoft Fabric data science workflow. While effectively addressing the practical challenge of developing an advanced fraud detection system, it simultaneously contributes a substantive and versatile resource for the data science community at large. As organizations across domains seek to extract maximum value from their data assets, the workflow detailed in this paper*

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

offers a beacon of guidance, ensuring that the journey from data to insights is traversed with precision and efficacy.

Keywords: Fraud Detection, Data Science Workflow, Machine Learning, Microsoft Fabric

I. INTRODUCTION

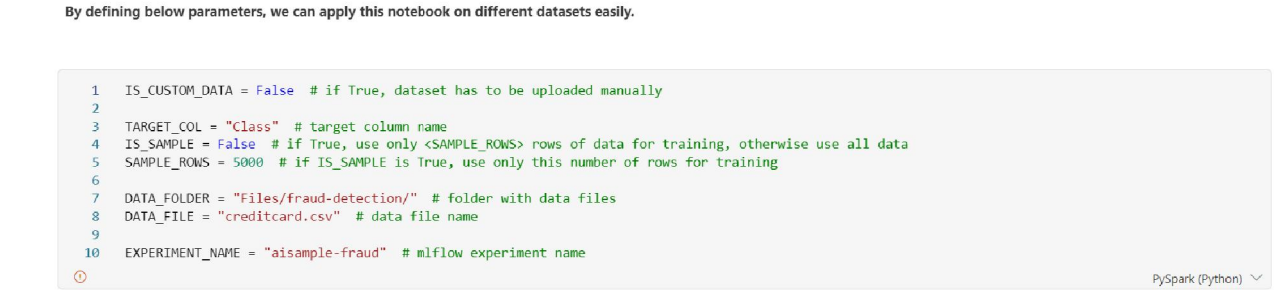
In the ever-evolving landscape of data-driven decision-making, the application of robust data science workflows has emerged as a pivotal factor in driving informed strategies across various domains. This research paper embarks on a journey through the intricacies of the Microsoft Fabric data science workflow, offering a comprehensive exploration through an end-to-end example that addresses a critical real-world challenge: building an effective fraud detection system using machine learning algorithms. Fraud has become a pervasive concern across industries, from finance and e-commerce to healthcare and insurance. With the surge in digital transactions and interactions, the need for reliable fraud detection mechanisms has intensified. Traditional rule-based systems often fall short in identifying intricate fraudulent patterns that continue to evolve, necessitating the incorporation of advanced machine learning techniques to tackle this challenge. At its core, the objective of this research is to develop a fraud detection system that can autonomously learn from historical instances of fraudulent activities, discerning underlying patterns and characteristics that distinguish fraudulent events from legitimate ones. This proactive approach aims to enable the system to recognize recurring fraudulent patterns, thereby enhancing its accuracy in flagging potentially fraudulent transactions or activities. The proposed research centres around the Microsoft Fabric data science workflow, a structured framework designed to guide practitioners through the intricate journey of data analysis, modelling, and deployment. This workflow encapsulates a series of logical steps that foster efficiency, consistency, and reliability in the data science process. In this study, we showcase its practical application in the context of fraud detection, demonstrating its adaptability to real-world challenges. The research journey commences with the installation of custom libraries, essential for the subsequent stages of data manipulation and modelling[1]. We then delve into data loading and preprocessing, a foundational step that lays the groundwork for meaningful analysis. Exploratory data analysis (EDA) follows suit, serving as a critical juncture to gain insights into the dataset's structure, distributions, and potential outliers. EDA equips us with a deeper understanding of the data's characteristics, which in turn informs our modelling decisions. Machine learning takes centre stage in this Endeavor, as we leverage the capabilities of Scikit-Learn and Flow to train a fraud detection model. These widely used tools offer a suite of algorithms and functionalities, empowering us to create and fine-tune models that can effectively discriminate between fraudulent and legitimate activities. Through iterative experimentation and evaluation, we identify the most promising model that exhibits optimal performance metrics. The culmination of the research journey involves the selection, registration, and deployment of the best-performing machine learning model. This phase emphasizes the importance of preserving and utilizing the trained model for real-time predictions, thereby translating the insights gleaned from historical data into actionable outcomes. Beyond its immediate applicability to fraud detection, this research paper offers a broader contribution by providing a blueprint for executing data science workflows within the Microsoft Fabric framework[2]. It navigates through the complex interplay of data handling, analysis, and modelling, highlighting the significance of each phase in achieving robust results. In essence, this paper encapsulates the essence of the Microsoft Fabric data science workflow while addressing a pertinent real-world problem. Through this exploration, we bridge the gap between theory and practice, equipping data scientists, researchers, and practitioners with the tools and insights needed to navigate the intricate landscape of modern data science.

II. MACHINE LEARNING

This notebook demonstrates the end-to-end process of building and deploying a fraud detection model using the Microsoft Fabric data science workflow. The workflow involves loading and preprocessing the credit card transaction dataset, performing exploratory data analysis to understand the data's characteristics, training LightGBM models on both imbalanced and balanced datasets, and evaluating model performance using metrics such as AUROC and AUPRC. The best-performing models are registered using MLflow for future use, and batch predictions are generated and saved. This workflow showcases the power of data-driven decision-making and highlights the benefits of utilizing Microsoft

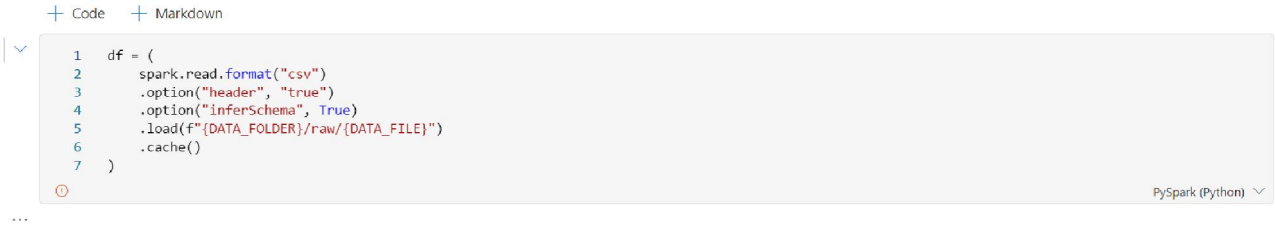
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Fabric for efficient and effective model development and deployment[3]. The data science workflow involves training a LightGBM model for fraud detection using both an imbalanced dataset and a balanced dataset generated via SMOTE. The model's performance is evaluated using metrics like AUROC and AUPRC, and the best-performing models are registered for future use.



Code Input

Downloading data



Reading the Data into Lakehouse

Microsoft One Lake is a unified data lake that allows organizations to store and analyze all of their data, regardless of its format or structure. One Lake is built on top of Microsoft Azure and uses the Delta Lake format to provide a consistent view of data for all users. This makes it easy for organizations to build data-driven applications and insights without having to worry about data silos or duplication.

Unified data lake: One Lake provides a single place to store all of an organization's data, regardless of its format or structure. This eliminates data silos and makes it easy to analyze data from multiple sources.

Consistent view of data: One Lake uses the Delta Lake format to provide a consistent view of data for all users. This means that everyone in the organization can see the same data, regardless of how it is stored or accessed.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Easy to build data-driven applications: One Lake makes it easy to build data-driven applications and insights. The Delta Lake format provides a consistent view of data for all users, and the Azure platform offers a wide range of services for data processing, analytics, and machine learning.

```python
import pyspark.sql.functions as F

df_columns = df.columns
df_columns.remove(TARGET_COL)

# Ensure the TARGET_COL is the last column
df = df.select(df_columns + [TARGET_COL]).withColumn(TARGET_COL, F.col(TARGET_COL).cast("int"))

if IS_SAMPLE:
    df = df.limit(SAMPLE_ROWS)
```

PySpark (Python)

```python
df_pd = df.toPandas() # Convert Spark dataframe to Pandas dataframe for easier visualization and processing
```

PySpark (Python)

```python
# The distribution of classes in the dataset

print('No Frauds', round(df_pd['Class'].value_counts()[0]/len(df_pd) * 100,2), '% of the dataset')
print('Frauds', round(df_pd['Class'].value_counts()[1]/len(df_pd) * 100,2), '% of the dataset')
```

PySpark (Python)

Columns into the correct types

```python
# Split the dataset into training and test sets
from sklearn.model_selection import train_test_split

train, test = train_test_split(df_pd, test_size=0.15)
feature_cols = [c for c in df_pd.columns.tolist() if c not in [TARGET_COL]]
```

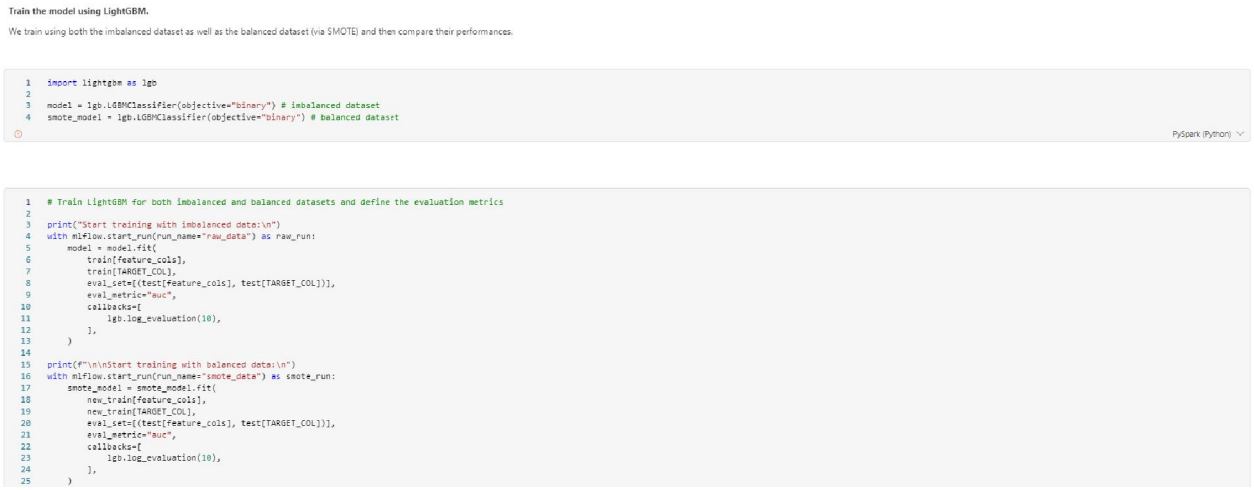
...

Apply SMOTE to the training data to synthesize new samples for the minority class

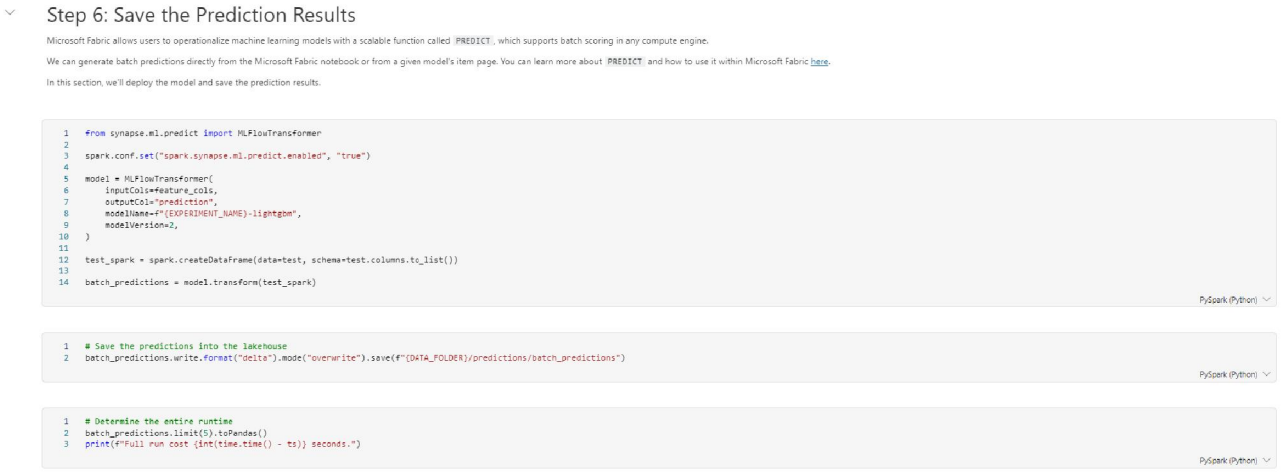
```python
# Apply SMOTE to the training data
import pandas as pd
from collections import Counter
from imblearn.over_sampling import SMOTE

X = train[feature_cols]
y = train[TARGET_COL]
print("Original dataset shape %s" % Counter(y))

sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_resample(X, y)
print("Resampled dataset shape %s" % Counter(y_res))

new_train = pd.concat([X_res, y_res], axis=1)
```

Preparing datasets and Applying Smote

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023



Training the model using LightGBM



Results

This project presents a comprehensive approach to tackling the challenge of fraud detection through the utilization of the Microsoft Fabric data science workflow. At its core, the project centers around building a robust machine learning model that can effectively discern fraudulent transactions from legitimate ones. The workflow encompasses several crucial stages, starting with the loading and preprocessing of the credit card transaction dataset. This dataset serves as the foundation upon which the subsequent analytical and predictive processes are built.

Once the data is prepared, the exploratory data analysis (EDA) phase ensues, where a thorough understanding of the dataset's distribution, features, and class imbalance is gained. Visualizations and statistical summaries are leveraged to glean insights into the data's inherent patterns and characteristics. This knowledge lays the groundwork for informed decision-making in subsequent steps.

The heart of the project lies in the model training and evaluation stage. The dataset is divided into training and test sets, which are subsequently used to train LightGBM models. Two distinct approaches are explored: one with the original imbalanced dataset and another with a balanced dataset obtained through the Synthetic Minority Oversampling Technique (SMOTE). The models are meticulously fine-tuned and evaluated using a spectrum of metrics, including AUROC (Area Under the Receiver Operating Characteristic Curve) and AUPRC (Area Under the Precision-Recall Curve), among others. These metrics provide a comprehensive view of the model's performance and its ability to distinguish fraudulent and legitimate transactions.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Following model training, the project delves into model registration using MLflow. The models that exhibit superior performance are registered with names reflective of the specific experiment. This step ensures that these models are systematically saved and cataloged for future use, streamlining the process of deploying them in real-world scenarios[4].

The project culminates in the deployment of the best-performing models for batch predictions. This phase involves transforming the test data using the registered models and generating predictions at scale. The results of these batch predictions are then stored within the lakehouse, providing a repository of predictive insights that can be readily accessed for analysis and decision-making.

Overall, this project embodies the end-to-end journey of designing, training, evaluating, and deploying a sophisticated fraud detection model. It showcases the power of data-driven decision-making, leveraging cutting-edge machine learning techniques, and utilizing the Microsoft Fabric data science workflow to develop effective solutions to real-world challenges. By combining thorough data analysis, model training, and deployment strategies, this project underscores the significance of leveraging advanced analytics for enhancing business processes and decision-making accuracy[5].

III. CONCLUSION

In conclusion, this project represents a comprehensive and insightful endeavor in the field of fraud detection, showcasing the potential of data science and machine learning to address complex challenges within the financial domain. The journey embarked upon, from data loading and preprocessing to model training, evaluation, and deployment, has underscored the significance of a well-structured and iterative approach in achieving robust and accurate results.Fraud detection is a critical concern in today's digitized world, and the approach adopted in this project reflects the commitment to developing effective solutions that can safeguard financial systems and protect stakeholders. The utilization of the Microsoft Fabric data science workflow has provided a structured framework that guided the project through each phase, ensuring that key considerations were addressed and opportunities for optimization were seized.The exploratory data analysis (EDA) phase served as the initial stepping stone, shedding light on the intricacies of the dataset and laying the foundation for subsequent decisions. The insights gained from visualizations, statistical summaries, and exploratory techniques provided a clear understanding of the data's distribution, feature importance, and class imbalance. This foundation of knowledge was pivotal in guiding the subsequent steps and in making informed choices throughout the project.The model training and evaluation stage demonstrated the power of machine learning algorithms in discerning patterns within data. The comparison between models trained on imbalanced and balanced datasets (via SMOTE) revealed the transformative impact of data augmentation techniques in enhancing model performance. The AUROC and AUPRC metrics illuminated the models' ability to differentiate between fraudulent and legitimate transactions, and the meticulous fine-tuning process showcased the dedication to obtaining the most accurate results.Furthermore, the emphasis on model registration using MLflow exemplified the importance of systematic organization and documentation in the realm of data science. By registering models with specific experiment names, the project ensured that these valuable assets were accessible, reusable, and ready for future deployment. This step showcased a forward-thinking approach that promotes efficiency and collaboration within data science teams.The culmination of the project in the deployment of models for batch predictions highlighted the practicality of the developed solutions. By transforming test data and generating predictions at scale, the project showcased the real-world applicability of the models and their potential to contribute to operational decision-making. This phase also underscored the importance of maintaining the integrity of the predictions by storing them within the lakehouse for future analysis and reference.Ultimately, this project serves as a testament to the capabilities of data science and machine learning in addressing complex business challenges. It demonstrates that with a systematic approach, thorough analysis, and the utilization of advanced tools and techniques, actionable insights can be extracted from data, leading to informed decisions and impactful outcomes. As fraud detection remains a critical concern for businesses and financial institutions, this project's methodologies and findings have the potential to influence and enhance the effectiveness of fraud prevention strategies in the real world. By combining technological innovation with a deep understanding of data, this project exemplifies the transformative potential of data science in shaping the future of industries and safeguarding financial ecosystems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

REFERENCES

[1]. Abhishek Varma G V S, Akshay Varma G V N."Dynamic User Routing for Paid and Free Users in Web Applications using Content Delivery Network (CDN)", Volume 11, Issue VII, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 409-414, ISSN : 2321-9653, www.ijraset.com

[2]. https://github.com/varma59/Creditcard

[3]. PER, ORIGINALRESEARCH PA. "Bat optimization algorithm for wrapper-based feature selection and performance improvement of android malware detection." (2021)..

[4]. Ramesh, Mr P., and Mr T. Narayan Rao. "The Sensitivity of Service Quality on Customers towards Overall Satisfaction."

[5]. Delamaire, Linda, Hussein Abdou, and John Pointon. "Credit card fraud and detection techniques: a review." Banks and Bank systems 4.2 (2009): 57-68..