**BUSINESS CASE STUDY**

**Q1. Data cleaning including missing values, outliers and multi-collinearity.**

Answer: - Handled missing values, as our dataset was having only 2-3 missing values in total, as compared to our large dataset. So, dropping them has been considered as a suitable option. Otherwise, we could have imputed the values using Sklearn’s, “Simple Imputer” or could have used other options such as forward fill or backward fill. But in our case that was not required. Have visualized the outliers using boxplot, but did not remove them, as our Machine Learning models such as Random Forest, and Decision Tree was capable enough to handle them. Have visualized the multi-collinearity, correlation values, using Heat Map. In which we understood the correlation among various attributes of the dataset. And found that ‘**oldbalanceOrg’ vs ‘newbalanceOrig’** and **‘oldbalanceDest’ vs ‘newbalanceDest’** has the highest correlation among all.

**Q2. Describe your fraud detection model in elaboration.**

Answer: - I have attached a “ReadMe” file to elaborate the model in detail. In that file, each and every step, from data loading to building prediction system has been elaborated with great detail. Just have a look at it.

**Q3. How did you select variables to be included in the model?**

Answer: - I have dropped two features from the dataset, which are ‘**nameOrig’** and ‘**nameDest’** as they are categorical features and also were least important for our model. Rest all features were selected. There was one more categorical feature with the name ‘type’, I used “**Label encoding**” to convert this categorical feature to numerical attribute. I encoded them as **['PAYMENT':0,'TRANSFER':1, 'CASH\_OUT':2, 'DEBIT':3, 'CASH\_IN':4]**, to convert them into numerical attribute. I did that because machine learning algorithms are good in handling numerical attributes as compared to categorical features.

**Q4. Demonstrate the performance of the model by using best set of tools.**

Answer: - I used confusion matrix, classification report, and accuracy score to display each model’s performance. In which Precision, recall, F1-score, support, and accuracy has been displayed for each model. Other than that, I build a predictive system, which takes in the input and predict whether the transaction is “fraud” or “legit”. So, mentioned techniques accurately highlights the model’s performance.

**Q5. What are the key factors that predicts the fraudulent customers?**

Answer:- Key factors which predicts the fraudulent customers are:- **‘step’, ‘type’, ‘amount’, ‘oldbalanceOrg’, ‘newbalanceOrig’, ‘oldbalanceDest’, ‘newbalanceDest’, ‘isFlaggedFraud’**. Our **predictive system** takes in these values, in which the value of ‘**type’** attribute should be entered in numerical form as we have **encoded** it in numerical form. Rest will get entered, as it is. Our model will first convert them into **NumPy** array, then the values would get **scaled** by going through ‘**standard scaler**’ then our **model** will get **implemented** and it will predict whether the **transaction** is **fraud** or **legit**.

**Q6. Do these factors make sense? If yes, How? If not, How not?**

Answer: - **Yes**, These factors are making sense. Because or model has achieved **100%** accuracy for **training** dataset and **99%** accuracy for **testing** dataset. Since our model is working fine. So, we can safely say that, Yes, our factors are making sense.

**Q7. What kind of prevention should be adopted while company update its infrastructure?**

Answer: - While updating the infrastructure, specific prevention measures are crucial to ensure:-

1. **Data integrity and Security: -** Ensure the security and integrity of the data, used to train and validate the machine learning model during the infrastructure update. Implement encryption, access controls, and data masking techniques to protect sensitive data from unauthorized access.
2. **Version control & Backup: -** Maintain version control of the machine learning model and associated datasets to track changes and ensure reproducibility during the update process. Implement backup procedures to create copies of the model and datasets, before initiating any updates to counter the risk of data loss or corruption.
3. **Testing & Validation: -** Ensure thorough testing and validation of the machine learning model on the updated infrastructure to ensure that it performs as expected.
4. **Model Retraining & Fine Tuning: -** Retraining the machine learning model should be considered. Updated data and features should be used to counter the changes in fraud detection pattern
5. **Monitoring & Maintenance:-**Key performance metrics such as accuracy, precision, recall, and false positive rates should be monitored, to ensure that the model remains effective in detecting fraudulent activities.
6. **Rollback Plan:-**Develop a rollback plan to revert to the previous version of the machine learning model or infrastructure configuration in case of unexpected issues post-update.
7. **Collaboration & Communication**: - Maintain collaboration between data scientists, IT teams, fraud analysts, and other stakeholders involved in the update process to coordinate efforts and address any concerns or issues that arise.
8. **Compliance & Regulations**: - Ensure that the updated infrastructure and machine learning model comply with relevant regulations, standards, and industry best practices for fraud detection and data privacy.

**Q8. Assuming these actions have been implemented, how would you determine if they work?**

Answer: - Here are some approaches to assess whether the prevention measures are working as intended.

1. **Performance Tracking: -** Analyse the changes in performance matrices to understand that whether there have been any improvement following the implementation of the performance measures.
2. **Anomaly Detection: -** Monitor the occurrence of fraud-related incidents and anomalies detected by the fraud detection systems post-update. Compare the frequency and severity of incidents detected before and after the update to evaluate the effectiveness of the prevention measures
3. **User Feedback and Reporting: -** Gather feedback from users, regarding their experience with the updated infrastructure, any challenges faced or areas for improvement during the post-update period.
4. **Comparative Analysis:-** Perform comparative analysis between the performance of the updated infrastructure and prevention measures
5. **Compliance & Audit Reviews: -** Conduct regular compliance reviews and audits to ensure that the updated infrastructure and prevention measures comply with relevant regulations, standards, and internal policies.
6. **Continuous Improvement & Adaptation: -** Implement a feedback loop to incorporate lessons learned from incidents, testing, and user feedback into ongoing improvement efforts.