Key insights from the key statistics output generated by analyzing absenteeism.csv dataset.

Key stats of data:

	count	mean	std	min	25%	50%	75%	max
Social drinker	740.0	0.567568	0.495749	0.0	0.0	1.0	1.0	1.0
Social smoker	740.0	0.072973	0.260268	0.0	0.0	0.0	0.0	1.0
Pet	740.0	0.745946	1.318258	0.0	0.0	0.0	1.0	8.0
Weight	740.0	79.035135	12.883211	56.0	69.0	83.0	89.0	108.0
Height	740.0	172.114865	6.034995	163.0	169.0	170.0	172.0	196.0
BMI	740.0	26.677027	4.285452	19.0	24.0	25.0	31.0	38.0
Absenteeism	740.0	6.924324	13.330998	0.0	2.0	3.0	8.0	120.0

1. Impact of Social Drinking on Absenteeism (in respect to BMI)

Insight: According to our analysis, social drinkers are more common, making up 56.8% of the sample, while non-social drinkers account for 43.2%, which means **420 out of 740** people **drink socially**. We can assume that it might have some impact on employees not showing up to the work.

Explanation: Our analysis shows that 56.8% of employees are social drinkers, which is significantly higher compared to employees with pets or those who are social smokers. This suggests that social drinking is a common habit among the workforce. With an average **BMI of 26.6%,** which falls in the **obese** category, it's reasonable to consider that social drinking could be contributing to absenteeism, as it may impact both health and work attendance.

```
Social Drinker Counts:
Social drinker
1    420
0    320
Name: count, dtype: int64
```

```
BMI Mode: 0    31
Name: BMI, dtype: int64
```

2. Low Frequency of Social Smoking

Insight: The majority of employees (92.7%) are non-social smokers, while only 7.3% are social smokers

Explanation: Since social smoking is relatively rare, it might not have a significant overall impact on absenteeism**. Only 54 out of 740** are social drinkers which must not have too much impact on the average of absenteeism.

```
Social Smoker Counts:
Social smoker
0    686
1     54
Name: count, dtype: int64
```

3. **Pet Ownership Distribution**

Insight: Most employees either do not own pets (62.2%) or own a few (0, 1, or 2 pets), with pet ownership dropping off significantly as the number of pets increases.

Explanation: The median and mode number of pets is 0, indicating a common trend of pet ownership being relatively low. Analysing how pet ownership correlates with absenteeism could be insightful, and according to us does not have much impact on the absenteeism. As some studies suggest that pet owners might experience lower stress levels and potentially fewer absences due to improved mental health.

```
Pet Ownership Counts:
Pet
0    460
1    138
2     96
4     32
8      8
5      6
Name: count, dtype: int64
```

```
Median of Pet: 0.0
Mode of Pet: 0    0
Name: Pet, dtype: int64
```

4. **Absenteeism Distribution and Outliers**

Insight: Absenteeism has a wide range (from 0 to 120 days), with a median of 3 days and a mode of 8 days. The large standard deviation (13.33) indicates high variability in absenteeism.

Explanation: The wide range and high standard deviation suggest that absenteeism varies greatly among employees. While the median absenteeism is relatively low (3 days), the high maximum value (120 days) and the mode (8 days) indicate that there are a few employees with significantly higher absenteeism. Investigating the causes for these extreme values and outliers could help in understanding and potentially mitigating excessive absenteeism.

```
Standard Deviation of Absenteeism: 13.330998100978201
Mean of Absenteeism: 6.924324324324324
Median of Absenteeism: 3.0
Mode of Absenteeism: 0    8
Name: Absenteeism, dtype: int64
```

Applying the CRISP-DM model to build a Fraudulent Detection System in banking

Why We Chose This Sector

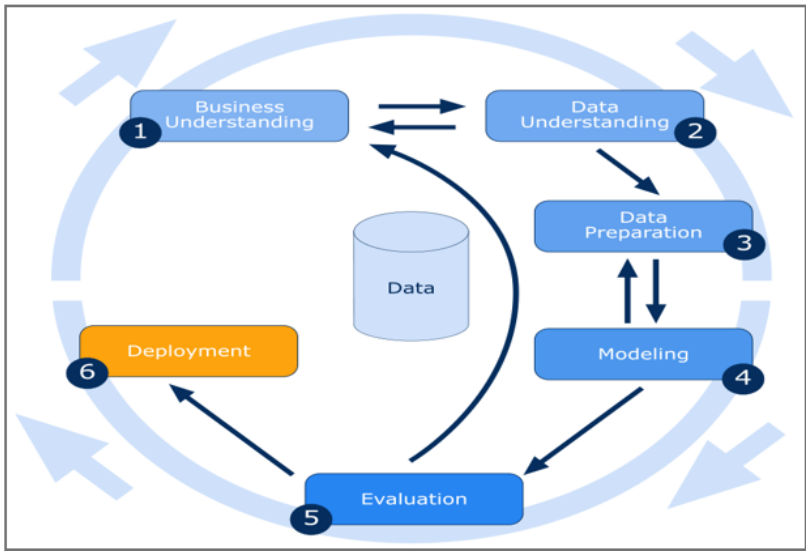
As a group of six Indian students who moved to Canada as international students, we have encountered numerous fraudulent calls and messages related to banking. This issue has caught our attention and motivated us to address it through data-driven solutions. We aim to develop a model to detect such fraudulent activities, starting by applying the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology.By focusing on the banking sector, we hope to mitigate the risks associated with fraudulent transactions and enhance security for individuals and institutions alike.

Introduction

Financial institutions are at risk of substantial financial losses and a loss of consumer trust because of fraudulent transactions. We can systematically create a fraud detectionalgorithm that identifies suspicious transactions in real time by utilizing the CRISP-DM (Cross Industry Standard Process for Data Mining) model. This assignment outlines the process of implementing the CRISP-DM methodology to develop a fraud detection system for financial services.

⇒ **The CRISPR-DM model majorly follows 6 phases that are as follows:**

1)	Business Understanding
2)	Data Understanding
3)	Data Preparation
4)	Modeling
5)	Evaluation
6)	Deployment



Phase 1: Business Understating

The objective is to establish a fraud detection system that promptly flags the bank to suspicious transactions, thereby facilitating the reduction of financial losses andthe improvement of customer confidence.

Goals:

1. **Detect fraud in real time**: The algorithm will detect suspect or fraudulent transactions as soon as they occur.
2. **Limit false positives to a minimum**: In order to prevent the inconvenience of legitimate clients, it is imperative to decrease the number of false positives (transactions that are incorrectly identified as fraudulent).

Success Criteria:

Success Criteria will be based on two parameters, the first being the ability of the model to detect at least 95% of the fraudulent transactions and the second being the limiting the false positive rate under 2%.

Phase 2: Data Understanding

During this phase, we will ascertain the data sources that are essential for the development of the fraud detection system and investigate the integrity of the data. Typically, banking transactions yield an abundance of information that is beneficial forthe detection of fraud.

Key Data Sources:

1. **Transaction Data**: Includes transaction amounts, timestamps, transaction types(e.g., withdrawals, purchases), and merchant information.
2. **Customer Data**: Details such as customer ID, account balance, historicaltransactions, and geographical location.
3. **External Data**: Geolocation, public reports of fraud, or historical fraud patternsacross other banks.

Initial Data Examination:

1. **Outlier Detection**: Identify unusual patterns in transaction amounts or behaviors that deviate from a customer's normal activities.
2. **Patterns in Historical Fraud Cases**: Examine past fraudulent transactions to identify patterns (e.g., sudden large withdrawals, unusual geographic locations).

Phase 3: Data Preparation

The data preparation process entails cleansing and converting the raw data to assure itsreadiness for analysis. This stage is essential for developing an effective fraud detection algorithm.

Steps Involves:

1. **Data Cleaning**: Remove duplicate transactions, handle missing or incomplete data entries, and standardize transaction fields (e.g., currency formats).
2. **Feature Engineering**: Create new features such as:
    - **Transaction Velocity**: The number of transactions in a short time span.
    - **Geographical Anomalies**: Transactions occurring in locations far from thecustomer's usual activities.
    - **Merchant Type**: Identify if the transaction involves a high-risk merchant(e.g., gambling or high-value luxury stores).
3. **Data Normalization**: Normalize numerical values such as transaction amounts to ensure consistency across different datasets.
4. **Splitting Data**: Split the dataset into training (80%) and testing (20%) subsets totrain the model and evaluate its effectiveness.

Phase 4: Modeling

The modeling phase involves selecting specific algorithms to predict suspicioustransaction are fraudulent or not.

Selected Modeling Techniques:

1. **Supervised Learning (Classification)**: Algorithms such as decision trees, random forests, or logistic regression can be used to classify transactions as fraudulent orlegitimate.
2. **Unsupervised Learning (Anomaly Detection)**: Techniques such as k-means clustering or isolation forests can identify transactions that deviate from normalpatterns, flagging them as potentially fraudulent.

Training and Tuning:

1. **Model Training:** Train the supervised models using historical transaction data, including labeled fraudulent and non-fraudulent transactions.
2. **Hyperparameter Tuning:** Adjust parameters in models (e.g., max depth for decision trees) to improve prediction accuracy and reduce false positives.
3. **Evaluation:** Test the models using the testing dataset to ensure they accurately detect fraud while minimizing false positives.

Phase 5: Evaluation

This phase involves evaluating the performance of the models to ensure they meet the business objectives.

Key Evaluation Metrics:

1. **Accuracy**: Measure how well the model correctly identifies both fraudulent and legitimate transactions.
2. **Precision and Recall:** Precision is the ratio of true frauds identified over the total flagged transactions, while recall measures the ability of the model to identify all fraudulent transactions.
3. **ROC Curve and AUC (Area Under the Curve):** These metrics will be used to balance the trade-off between true positives and false positives, ensuring that the model is neither too lenient nor too aggressive.

Model Refinement:

1. If the model underperforms (e.g., high false positive rate), I would return to the data preparation or modeling phases to fine-tune the features or algorithms.
2. For instance, if high-value legitimate transactions are often flagged as fraudulent, I might introduce additional customer-level features, such as purchase history, to distinguish these patterns.

Phase 6: Deployment

After evaluating and refining the model, the next step is to deploy it in the bank's transaction processing system to identify fraud in real time.

Steps for Deployment:

1. **Real-Time Integration:** Integrate the fraud detection system into the bank's existing infrastructure to monitor transactions as they occur.
2. **Alert System:** Create an automated alert system that flags suspicious transactions for further review by fraud analysts.
3. **Customer Interaction:** Implement a protocol for customer notifications when transactions are flagged for review, minimizing inconvenience to legitimate customers.

Monitoring and Maintenance:

1. **Model Monitoring:** Continuously monitor the model's performance in real-world environments and ensure it adapts to evolving fraud tactics.
2. **Regular Updates**: Periodically retrain the model with new data to capture emerging fraud patterns or customer behavior changes.

Conclusion

By following the CRISP-DM methodology, we can develop a robust fraud detection system that accurately flags suspicious transactions in real time. This structured approach ensures the model is aligned with business objectives, built on high-quality data, and adaptable to evolving threats. Ultimately, this system will enhance fraud prevention efforts and safeguard both the bank and its customers from potential financial harm.