# Credit Card Fraud Detection Project Overview

## Project Overview: Understanding the Basics and Goals

The goal of this project is to build a robust machine learning model capable of detecting fraudulent credit card transactions. Credit card fraud is a growing concern in today's increasingly digital financial landscape, leading to significant financial losses for both consumers and financial institutions.  
  
This project focuses on analyzing a real-world dataset of credit card transactions in the US to uncover patterns indicative of fraudulent activity. Through:  
  
- 📊 Exploratory Data Analysis  
- 🛠️ Feature Engineering  
- 🤖 Machine Learning Modeling

● What are we trying to find out?  
We aim to identify whether a transaction is fraudulent based on transaction metadata and user behavior.

● What do we already know?  
We know that fraud is rare compared to legitimate transactions, approx 0.5% of credit card transactions are fraudulent making this a highly imbalanced classification challenge. Fraudulent behavior often exhibits patterns like abnormal transaction times, high-value amounts, geographic anomalies and combinations between them e.g. a senior customer is less likely to perform transactions at certain amounts and times of the day.

● What are we aiming to achieve?  
Our goal is to accurately detect fraudulent transactions while minimizing false positives to avoid disrupting genuine users.

● What factors affect our results?  
Class imbalance, feature quality, model selection, model parameters, and evaluation metrics all significantly influence the final model's effectiveness.

● Is there something new we can use?  
Yes. We explore advanced ensemble models like XGBoost, feature interaction engineering, and resampling strategies (SMOTE, SMOTETomek) to counter imbalance for the end goal os enhancing prediction accuracy and generalization.

## Project Stages Summary

1. Data Preparation:

* Started with an 11GB dataset and a relatively small customers csv, need to sample a large dataset (dask ) and verify the 300K rows sample and the dataset distributions are the same, done by both visually inspecting the fraud distribution over months by conducting Chi-square test, resulted p-value 0.655 (>> 0.05) indicates that both the sampled and larger data are the same, the selected model data is a 300K rows sampled from year 2020, our baseline will be **sample\_300k\_2020.csv.**
* Merged the customers.csv into sample\_300k\_2020.csv.
* Checked for NAN & missing values, none found
* Converted Datetime columns ot Datetime type.
* Converted numerical types to proper types.
* Converted string to proper types plus handled case, white spaces & special characters.
* Drop duplicates.
* Ensure proper ranges (geographical lat/long)
* Narrow *state* into 4 regions.
* From transaction date & time extract day\_of\_the\_week, hour, age.
* Narrow age into 6 age groups, child, teenager, young\_adult, middle\_aged, senior, elderly.
* Narrow transaction hour into 4 time segments, morning, afternoon, evening & late night.
* Narrow credit card number (first 4 digits of the cc\_num) into credit card service providers e.g. Visa, Mastercard, Discover, etc.
* Extract area category urban|rural from profile
* Narrow job high cardinality to 17 job definitions e.g. Finance, Engineering, healthcare etc.
* Compute and add the distance between the customer (lat/long) and the merchant (lat/long).
* Computed time between transactions from unix time.

Exported the prepared data to **sample\_300k\_2020\_prepared.pkl** for further use.

1. **EDA – Exploratory Data Analysis:** Visualized feature distributions, correlations, and fraud patterns over time and geography.

**Categorical Features EDA** running Chi-Square test *['category', 'region', 'job\_cat', 'cc\_type', 'area\_cat', 'trans\_time\_segment', 'age\_group']* against the target value *is\_fraud* found the following:

1. Data Cleansing: Handled missing values and removed/capped outliers in key numeric features like amount and distance.
2. One-hot Encoding: Encoded categorical variables using one-hot and label encoding based on cardinality.
3. Feature Engineering: Created new variables including time-distance interactions, log transformations, and categorical pairings.
4. Imbalanced Data Handling: Applied ROS, RUS, SMOTE, and SMOTETomek to improve recall while maintaining precision.
5. Model Selection and Fine-tuning: Tuned models like Random Forest, XGBoost, and AdaBoost using GridSearch and RandomizedSearchCV.

## Project Summary: Deployment and Beneficiaries of Machine Learning

● How will we deploy the Machine Learning?  
The trained fraud detection model can be deployed as an API service integrated into payment processing systems. It can evaluate transactions in real-time and flag high-risk activities for further investigation or automatic action.

● Who will use and benefit from the Machine Learning?  
Fraud analysts, banks, credit card companies, and end users benefit directly. It enhances fraud prevention efforts, reduces financial losses, and improves trust in digital transactions.