**Capstone Two - Project Ideas**

# Project Idea -1 “House Price Prediction”

# **Problem Statement**

* The data source for this project is a CSV file containing real estate listings from various regions in the US. For many people, owning a home is a big deal not only financially, but in the whole life journey. Finding a house that fits your needs can be difficult, with price being a major factor to consider.
* The objective of this project is to create a user-friendly tool that provides a preliminary estimate of house prices based on readily available information. While it may not guarantee pinpoint accuracy, it aims to offer homebuyers a helpful starting point for initiating price negotiations

**Content:**  The main data source for this project is a CSV file that contains real estate listings in different areas of New York, US. It contains 10 columns shown as below:

* status: house status, either for sale or ready to build
* bed: number of bedrooms in the house
* bath: number of bathrooms in the house
* acre\_lot: property/land size in acres
* city: City the house residents in
* state: State the house residents in
* zip\_code: postal code of the area
* house\_size: house size in square feet
* prev\_sold\_date: the previously sold date if it is recently sold
* price: house prices (either current listing price or recently sold price)

**Source:** [Dataset.csv](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset?select=realtor-data.zip.csv)

# Project Idea -2 “Wase User Churned”

**Problem Statement**

* In this Project we would be trying to understand the Waze user churn or retain dataset and also build models that would help us predict the possibilities from this dataset.
* Developing a churn prediction model will help prevent churn, improve user retention, and grow Waze’s business. Analyze and interpret data, generate actionable insights, and help leadership make informed business decisions.
* Proactively identify factors that will engage high-risk churn users to retain them.

**Content:**-The primary data source for this project consists of information stored in a CSV file.It exhibits 13 columns as outlined below:

* ID: Identification Number
* Lable: Binary target variable (“retained” vs “churned”)
* Sessions: The number of occurrence of a user opening the app during the month
* Drives: An occurrence of driving at least 1 km during the month
* total\_sessions: A model estimate of the total number of sessions since a user has
* n\_days\_after\_onboarding: The number of days since a user signed up for the app
* total\_navigations\_fav1: Total navigations since onboarding to the user’s favorite place 1
* total\_navigations\_fav2: Total navigations since onboarding to the user’s favorite place 2
* driven\_km\_drives:Total kilometers driven during the month
* duration\_minutes\_drives: Total duration driven in minutes during the month
* activity\_days: Number of days the user opens the app during the month
* driving\_days: Number of days the user drives (at least 1 km) during the month
* device: The type of device a user starts a session with

**Source:** [Dataset.csv](https://www.kaggle.com/datasets/juliasuzuki/waze-dataset-to-predict-user-churn?select=waze_dataset.csv)

# Project Idea -3 “Vehicle Insurance Fraud Detection”

# **Problem Statement**

* Apparently, it's a real database of an American insurance company Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident. Our objective is to classify claims as fraud or legit.
* Some common examples include staged accidents where fraudsters deliberately “arrange” for accidents to occur; the use of phantom passengers where people who were not even at the scene of the accident claim to have suffered grievous injury, and make false personal injury claims where personal injuries are grossly exaggerated.

**Content:-** This dataset contains 33 columns and 15420 rows like vehicle dataset - attribute, model, accident details, etc along with policy details - policy type, tenure, etc. The target is to detect if a claim application is fraudulent or not - FraudFound\_P

Source: [Dataset.csv](https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection)