**CREDIT CARD FRAUD DETECTION**

Data Source: <https://www.kaggle.com/datasets/kartik2112/fraud-detection?resource=download>

We considered used car data and performed some basic data cleaning. We dropped nearly 2000 instances(data points) based on number of null values it had. We dropped all instances that had more than 4 null values. To handle null values in transmission and condition. We grouped cars based on manufacturer and body type and then imputed average values of each group to the missing data for the group it belongs. We used custom function for the imputation purpose. We performed other basic data preprocessing.  
Once basic preprocessing is completed, we performed rigorous EDA. We checked how target variable vehicle price is dependent on other variables. We checked how manufacturer, body type, vehicle color, condition etc.. impacts. Following is quick summary of what we found:

* The resale value was high for luxury car brands. The resale price was in low to medium range for niche brands like Ford, Chevrolet, Toyota etc..
* Convertible, Coupe, SUV and some special editions( which were marked as other) increased price of car, whereas body type like Sedan, mini-van, hatchback, wagon were relatively less priced comparing their counterpart
* Automatic Transmission vehicles were found to be expensive than manual
* With vehicle condition deteriorating, the price dropped significantly. But for luxury cars the price was still high even if condition was average.
* Black, White, Silver, Gray, Blue and Red were popular car colours and were high in numbers. So we grouped all other colours together as Other
* For interior people preferred black, tan, gray and beige colors.
* Some luxury cars had some unique colors which affected average price during EDA
* Price depreciated with age of vechiles, but depreciation was high for affordable cars than luxurious cars.
* Odometer reading played a keyrole in determining price. And there was some correlation between odometer and age of car which is obvious but correlation was not too high( around -0.5)

Then we checked for statistical significance of each feature. We checked for skewness, correlation etc.. Though there was slight correlation among some features after encoding and log + Standard Scaler transformation. most of correlation was removed. We checked p value and VIF for each feature. We found there was very less correlation thus confirming independence of each features.

We built pipeline to perform categorical encoding and Scaling Transformation and model fitting. We used scikileanr's Pipeline module.

For modelling we tried with Linear Regression, Ridge, Lasso and Random Forest Regression. All of the four performed quite good. But RandomForest was most suitable amongst all. So we went ahead with Random Forest for Hyperparameter Tuning.

We got best parameter at n\_estimators': 150, min\_samples\_split: 10, min\_samples\_leaf: 2, max\_depth: 10  
We obtained following evaluation metrics:  
Note the MAE error is on log value(since we did log scaling).

𝑇𝑟𝑎𝑖𝑛 𝑆𝑒𝑡::𝑅2=0.949 𝑎𝑛𝑑 𝑀𝐴𝐸=0.115

𝑇𝑒𝑠𝑡 𝑆𝑒𝑡::𝑅2=0.946 𝑎𝑛𝑑 𝑀𝐴𝐸=0.1185

MAE on car price is around 950$