# ML ASSIGNMENT-2 (HACKATHON)

- GROUP 3 : Sarthak Khoche;Ayush Yadav

- TEAM MEMBERS: RAKSHIT GUPTA (IMT2019516) , RONIT JAIN (IMT2019073)

# ABOUT THE DATASET

- REGRESSION : With the help of two features *x* and *y*, predict the target variable *Target* using techniques of regression

- CLASSIFICATION: This dataset represents the activity of customers who are interested in buying insurance policies. The dataset has 300 features related to sales information, coverage information, personal information (of the customer), property information (for a home insurance) and geographic information. Task is to predict if a customer will buy a policy or not ( i.e., predict Conversion_result).

# APPROACH FOR REGRESSION TASK

- **Preprocessing** : There were neither NAN values nor duplicate rows in the dataset. The dataset consists of only numerical data thus, no LABEL ENCODING was required.

  We used StandardScaler() to standardize the features.

- **Model** : We applied three different models on the dataset . Following are the models we used :

  1) Linear Regression 2) Polynomial Regression 3) SVM.

  After comparing their predictions using MSE, Polynomial Regression gave us the least MSE.

# APPROACH FOR REGRESSION TASK

- **Hyperparameters** : We got best results in Polynomial Regression for degree = 31. In SVM , we tried different parameters like gamma, C, max_iter, etc. and compared their results.
- **Results** : MSE for training dataset : 3.90842

  Kaggle Score(MSE) : 3.78790(Public) , 4.20740(Private)

# APPROACH FOR CLASSIFICATION TASK

- **Preprocessing** : We did data cleaning of both training and testing data separately. For both training and testing data , the features with more than 75% missing data were dropped, also two features with only one category of data were dropped, and for features with more than 100 NAN values were filled using mean/median/mode method.

In training data, total of 9 features had NaN values, out of which 6 of them had less than 100 NAN values and their rows were dropped while in testing data NAN values of these 6 features were filled using mean/median/mode method.

There were no duplicate rows in the dataset. Also categorical data was converted into numerical data using Label Encoding. We used StandardScaler() to standardize the features.

# APPROACH FOR CLASSIFICATION TASK

- **Reducing the features**: In order to reduce 300 features, we used PCA and correlation separately to make predictions and compared results for both.
- **Model**: We applied three different models on the dataset . Following are the models we used :

  1) Logistic Regression 2) Naïve Bayes Classifier 3) SVM.

  After comparing their predictions using f1 score , Logistic Regression gave us the best result.
- **Hyperparameters** : In logistic regression, we tried different parameters like solver('*newton-cg*', '*saga*') penalty, random_state, max_iter. In SVM , we tried different parameters like gamma, C, max_iter, etc. and compared their results.
- **Results** : f1 score for training data: ***0.84762***

  Kaggle score(f1 score) : ***0.84891*** (public) , ***0.84923*** (private)