**DATA VERSE**

**Vehicle Insurance**

1. Adversarial validation - I was able to determine whether there is a feature that separates the train and test data, and I discovered that there isn't.

2. Feature Transformation –

* I transformed the necessary features by using calculated metrics for the vehicles' maximum torque and mac power.
* By removing the M and C characters, the area and model are converted to integers.
* To convert Yes/No features, ordinal mapping encoding is used.
* For nominal features, label encoding is used.

3. Feature selection - Using the chi-square test, I was able to eliminate a large amount of categorical data that were not significant to the claim.

4. Catboost model with 5 stratified kflold is used for machine learning.

5. Addressing Imbalance - I was able to use scale pos weight (10) by taking the average of np. sqrt(count(negative class)/count(positive class) + (count(negative class)/count(positive class) class)

6. Submission - setting a threshold and selecting a mode for the prediction improves performance on the leaderboard.