**BV0**

We designed a baseline with manual exploration

* Cross-Validation
* Generate Dummies (training join with validation)

Parameters Otimization

* Gridsearch

Pre-Processing

* Drop missing values (training, validation)
* Boxplot outlier removal percent= 0.03, 1.5(training)
* Normalize (Min-max Scaler)

Feature Engineering

* Extract Business features

**BV1**

* Cross-Validation (Stratify k-folds)
* Generate Dummies (training join with validation)

We choose stratified because it splits the dataset in a way that is well distributed, and consistent with the original distribution.

Parameters Otimization

* Gridsearch

Pre-Processing

* Impute missing values
* Mahalanobis
* Generate Dummies
* Normalize (Min-max Scaler)
* Balancing do dataset (weights) - SmoteNC

Undersampling vs Oversampling vs Smot vs Weight balancing vs Adesine

We decided to select between SMOT and ADESINE because our data was too small for undersampling. Oversampling is a simpler version of SMOT

[https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/](https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/?fbclid=IwAR3rieIqzjnvo-J3LFH0eTGdb6JTUb9X8K3SbR7gMkkECTk-O3ApRgtpqhQ)

Feature Engineering

* Extract Business features
* PCA

Feature Selection

* Correlation based feature selection (Feature importance based on correlation to Response)

**BV2**

Parameters Otimization

* Gridsearch

Pre-Processing

* Balancing do dataset (weights)
* Impute missing values
* Outlier Smoothing
* Generate Dummies
* Normalize (Min-Max Scaler)

Feature Engineering

* Extract Business Features
* Pca vs Lda vs Ica vs Factor Analysis (Look for comparisons on the internet)

Feature Selection

* Ranking feature selection

Missing:

* Bayesian Optimization

To-Do:

Correlations