INITIAL ACTIVITIES

Extract distributions from dictionary

Convert the dictionary with correlation matrix to dataframe

DATA GENERATION (DATA SYNTESIS)

START

Categorical Data
(Kolmogorov-Smirnof distribution test)

Numerical Data
(Copula)

Combine the two datasets

Apply the business rules

Check if the distributions align

No

Yes

FEATURE ENGINEERING

Procedures on Categorical features

One-hot encoding

Procedures on Numerical features

Rescaling
1. Z-score standartization

Non-linear transformations
1. Logarithmic
2. Quadratic
3. Reciprocal
4. Exponential
5. Square rooted

Precedures for generating custom features

Combine two features

FEATURE SELECTION

Categorical features

Chi Squared

Numerical features

ANOVA

Custom features

Domain knowledge and proof of visual interaction effects

MODELING

1. Logistic Regression
2. Random Forest
3. Decision Tree
4. XgBoost