TUNISIAN TAXES FRAUD

REPORT

abdelrahman seddik 2210000775

omar tamer 2210002828

mohamed nader 221000517

mohamed badea 221000627

1		Zen	5.123588415
2		MICADEE LAHASCOM	5.142392426
3		MEHBY Higher school of communication of tunis	5.145114236
4		ADILKhan17	5.158252834
5		TopOneSoon	5.169508233
6		☝️ 🤭 ☝️ Team	5.169900326

introduction

THERE IS A TAX FRAUD DETECTION SO WE TRY TO DETECT AND SOLVE TAX FRAUD USING SUPERVISD MACHINE LEARNING AND MODEL TRAINING AND ADVANCED FEATURE ENGINEERING

MOTIVATION

:IMPROVING THE ACCURACY OF TAX FRAUD DETECTION TO HRLP GOVERNMENT TO REDUCE MONEY LOSS

OBJECTIVES

: 1-HANDLING MISSING VALUE AND OUTLIERS

2-MAKE ACCURATE PREDICTIONS THAT IS MEASURED BY ROOTMEANEDSQUARE RMSE

3-CREATING ENGINERING FEATURES

4-REACH THE BEST RSME SCORE

3-EVALUATION MATRICS

WE ARE EVALUATING IT IS GOOD OR NO BY RMSE
THE LOWER IS THE BETTER

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

3-DATASET STRUCTURE

1-TRAIN.CSV: CONTAINS ID AND TARGET AND THE VALUE OF IT IS 15000

2-TEST.CSV: CONTAINS ID BUT NO TARGET AND THE VALUE OF IT IS 5000

3-SUMBISSION.CSV: IT CONTAIN A SAMPLE THAT MUST SUBMIT LIKE IT

DATA PREPROCCESING

1- LOAD DATA FROM TRAIN AND TEST
2-CHECK MISSING VALUES
3-SEE IT IS NUMERICAL OR CATEGORICAL
4-HANDLING MISSING VALUES
5-FEATURE ENGINEERING
6- CONVERT CATEGORICAL DATA INTO NUMERICAL
7-EDA

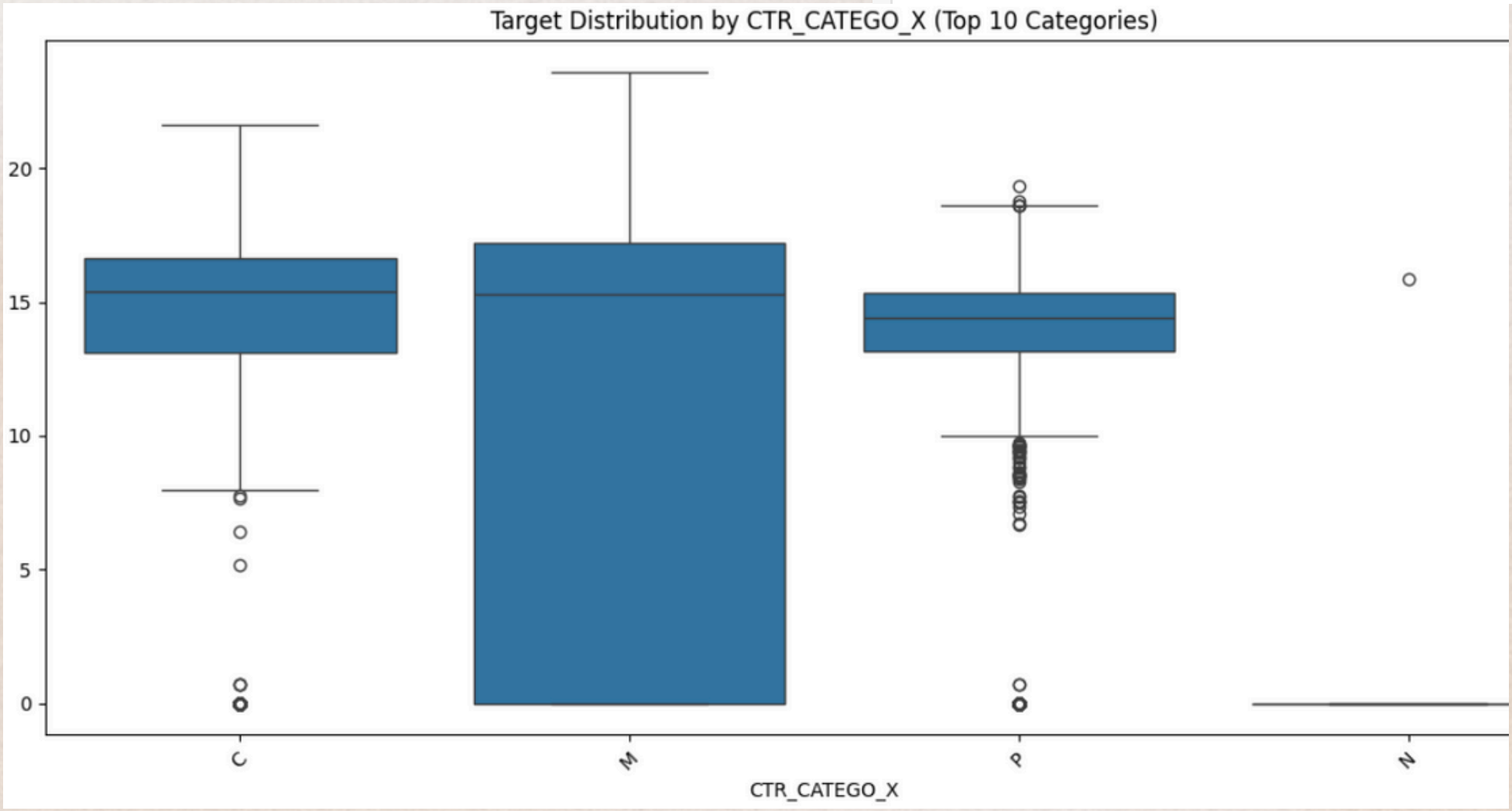
1-handle outliers by IQr

CALCALUATE IQR BY
CALCULATING Q3 AND Q1 AND
MINUS THEM TO GET THE IQR
THEN GET THE LOWER AND UPPER
BOUND
328 OUTLIER DETECTED

```
BEFORE outlier handling:
Mean: 11.7801, Std: 7.0858
Range: [0.0000, 23.5913]

Outlier bounds: [-24.9384, 41.5
Number of outliers detected: 0

AFTER outlier handling:
Mean: 11.7801, Std: 7.0858
Range: [0.0000, 23.5913]
```



missing values and class imbalance

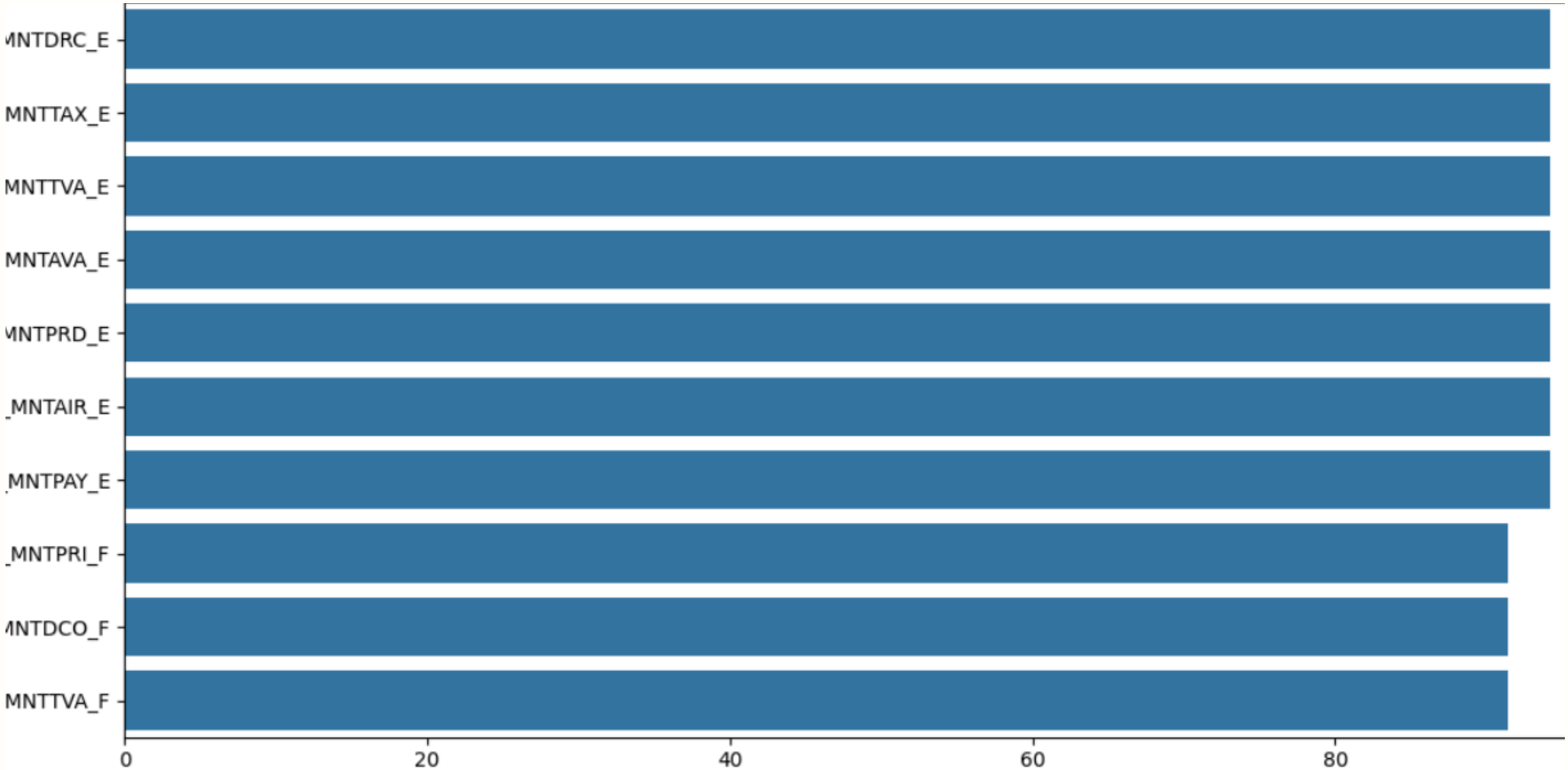
MISSING VALUES

- REPLACE MISSING VALUES WITH -999 BY FILLNA(-999)

-WHY -999? BECAUSE IT IS A LARGE OUTLIER SO YOU KNOW IT ARE OUTLIER

CLASS IMBALANCE

NO CLASS IMBALANCE BECAUSE ITS A REGRESSION MODEL


MNTDRC_E
MNTTAX_E
MNTTVA_E
MNTAVA_E
MNTPRD_E
_MNTAIR_E
_MNTPAY_E
_MNTPRI_F
MNTDCO_F
MNTTVA_F
0
20
40
60
80

target encoding and feature engineering

TARGET ENCODING

1- CALCALUTE MEAN

2-SPLIT DATA BY K FOLDS TO AVOID DATA LEAKAGE

3-NOISE ADDITION TO REDUCE OVERFITTING

4-CONVERT CATEGORICAL INTO NUMERICAL

```
1. Creating categorical feature interactions...
  - Created count and nunique features for CTR_MATFIS
  - Created count and nunique features for ACT_CODACT
  - Created count and nunique features for BCT_CODBUR
  - Created count and nunique features for CTR_CATEGO_X
  - Created count and nunique features for FJU_CODFJU

2. Creating financial ratio features...
  Processing base column: FAC_MNTPRI_F
    - Created 3 ratio features for FAC_MFODEC_F
    - Created 3 ratio features for FAC_MNTDCO_F
    - Created 3 ratio features for FAC_MNTTVA_F
  Processing base column: FAC_MNTPRI_C
    - Created 3 ratio features for FAC_MFODEC_C
    - Created 3 ratio features for FAC_MNTDCO_C
    - Created 3 ratio features for FAC_MNTTVA_C
  Total ratio features created: 18

3. Creating categorical feature pair interactions...
  - Created interaction: BCT_CODBUR_CTR_CATEGO_X
  - Created interaction: ACT_CODACT_FJU_CODFJU
  - Created interaction: CTR_MATFIS_ACT_CODACT
  - Created interaction: CTR_MATFIS_BCT_CODBUR
  - Created interaction: CTR_CATEGO_X_FJU_CODFJU
  Total interaction features created: 5

4. Creating numerical feature aggregations...
  - Created 6 numerical aggregation features
  - Aggregated across 152 numerical columns

5. Creating missing value indicators...
  - Created 122 missing value indicator features
```

FEATURE ENGINEERING

:1 · RATIOS AND LOG RATIOS AND AND DIFFERENCES BETWEEN RELATED COLUMNS

2· INTERACTION FEATURES TO COMBINE OF CATEGORICAL VARIABLES

3· INDICATE MISSING VALUES

· AGGREGATED STASTICS LIKE MEAN AND STD AND STD AND MAX AND MIN AND RANGE

4-

MODEL USED :LIGHT GBM

USES LIGHT GBM,A FAST AND EFFICIENT GRADIENT FOR FASTING FRAMEWORK

APPLY K FOLD CROSS VALIDATION TO TRAIN AND EVALUATE MODELS ON DIFFERENT SPLITS.

APPLY CATBOOSTENCODER ON CATEGORICAL VARIABLES WITHIN EACH FOLD FOR BETTER ENCODING

MONITOR VALIDATION RMSE AND USES EARLY STOPPING TO PREVENT OVERFITTING

function used

1-LOAD_AND_EXPLORE_DATA():
LOAD THE TRAINING AND TEST DATASETS

2-HANDLE_OUTLIERS()
DETECT AND HANDLE OUTLIERS IN THE TARGET

3-PERFORM_EXPLORATORY_ANALYSIS()
EXPLORE DATA DISTRIBUTION AND CORRELATIONS, AND MISSING VALUES

4-PREPARE_DATASETS_FOR_MODELING()
COMBINE TRAIN OR TEST FOR PREPROCESSING

5-CREATE_ADVANCED_FEATURES()
GENERATE NEW FEATURES FOR BETTER MODEL PERFORMANCE

6-TRAIN_LIGHTGBM_MODEL()
TRAIN A LIGHTGBM MODEL WITH CROSS VALIDATION

7-COMPARE_RESULTS()
EVALUATE MODEL PERFORMANCE VS BASELINE

8-GENERATE_SUBMISSION()
CREATE A READY SUBMIT FILE TO SUMBIT IT

RESULTS

RESULTS BEFORE:

1-THERE WAS MISSING VALLUES

2-THERE WAS OUTLIERS AND EXTREME VALUES

3- THERE WAS NO ENGINEERING FEATURES

4- THE RSME WAS 7.0856060

RESULTS AFTER:

1- HANDLED MISSING VALUES

2-HANDLED OUTLIERS

3-ADDED RATIOS AND ENGINERING FEAATURES

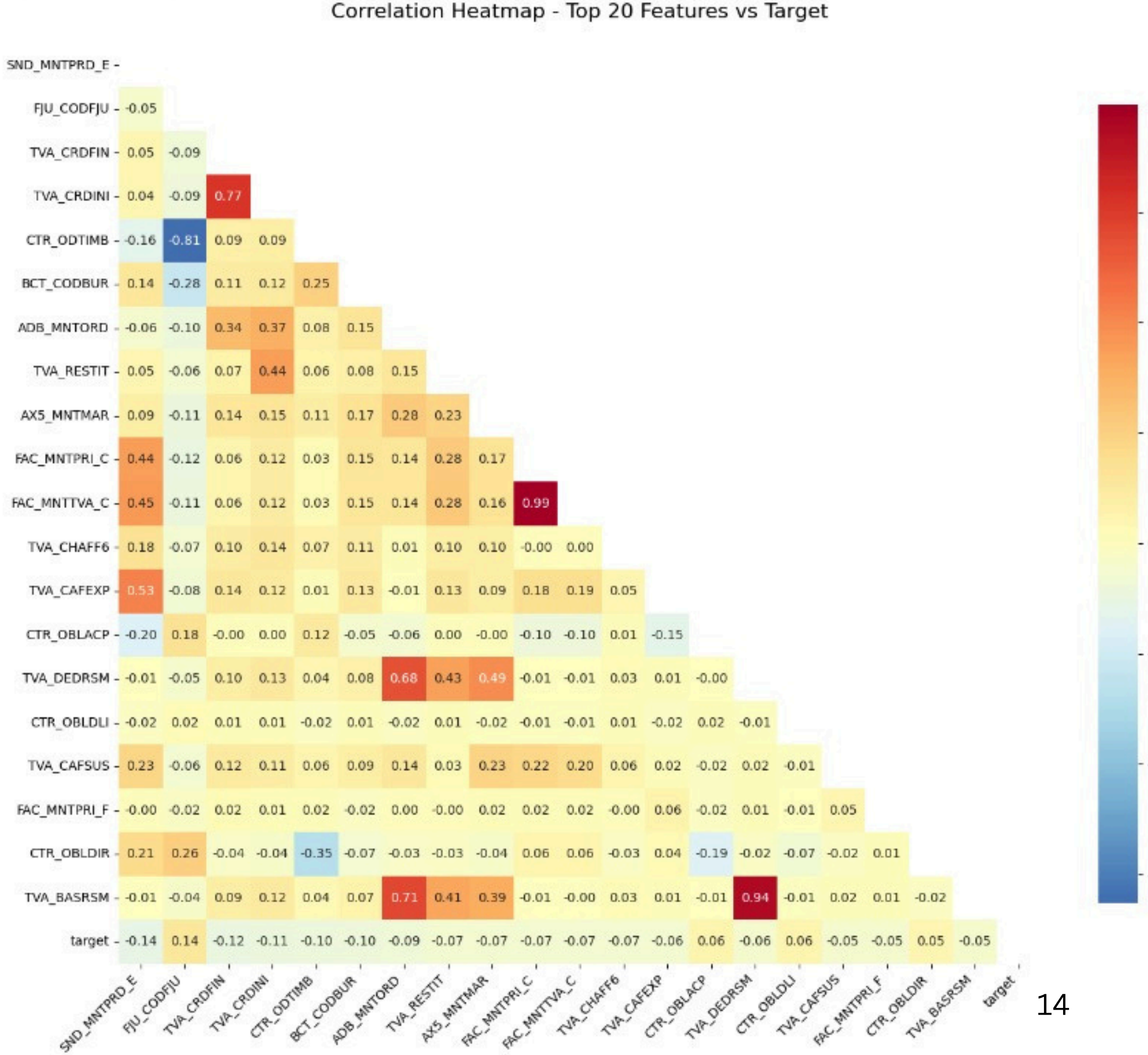
4-MODEL PERFORMA6.16NCE IS BETTER

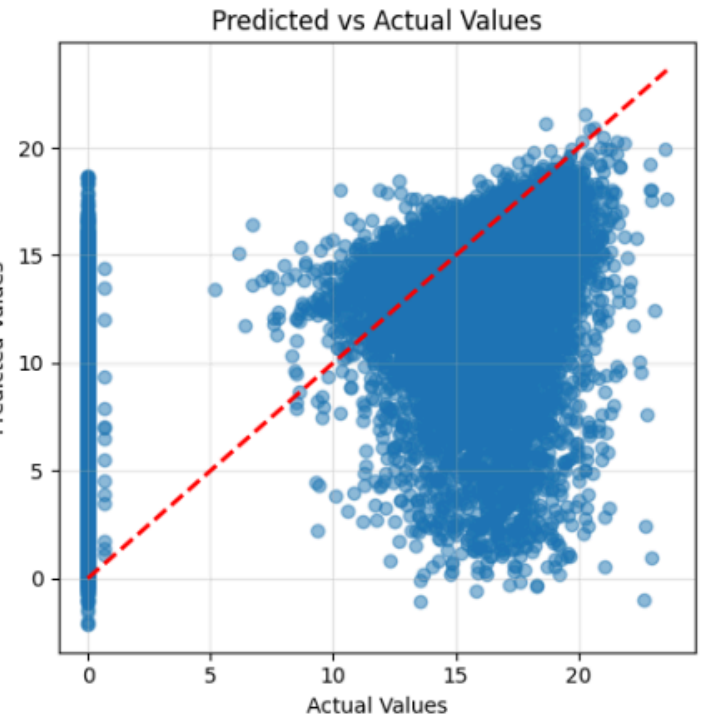
5- RSNE SCORE ARE 5.377

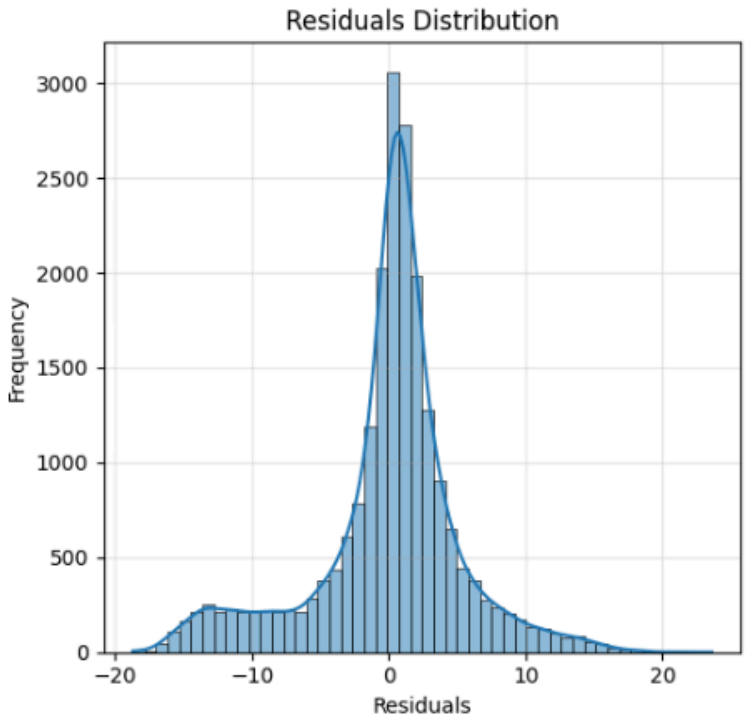
6- ENCODE CATEGORICAL VARIABLES WITHOUT LEAKAGE

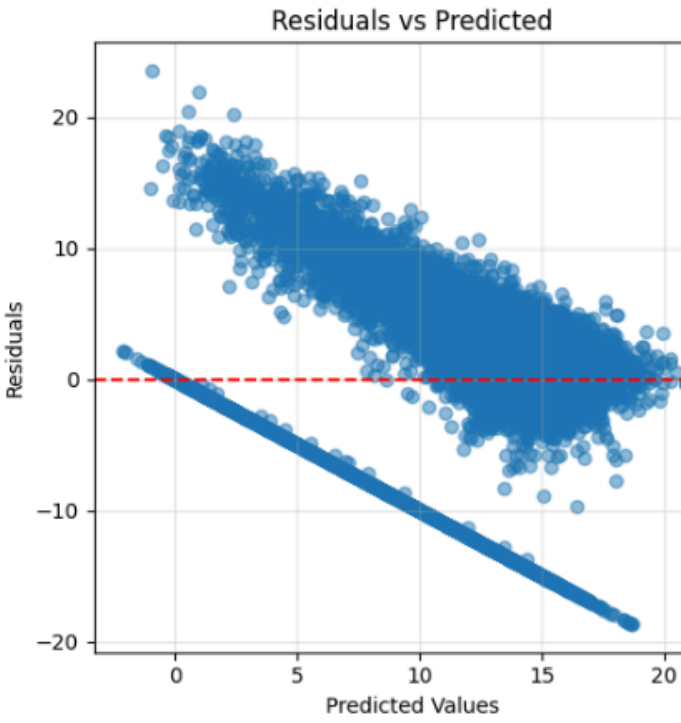
```
PERFORMANCE COMPARISON:
==========================================
Baseline RMSE (using mean): 7.08561
Model RMSE (after preprocessing): 5.37810
Improvement: 1.70750
Improvement %: 24.10%
```

Generating correlation heatmap...




Predicted vs Actual Values
Actual Values
Predicted Values
0
5
10
15
20


Residuals Distribution
Residuals
Frequency
0
500
1000
1500
2000
2500
3000
−20
−10
0
10
20


Residuals vs Predicted
Predicted Values
Residuals
−20
−10
0
10
20
0
5
10
15
20