Features Selection Roadmap

Legend:

◇

Events

▭

Dataframe

▯ File

---

model.csv

50000 x 19

◇ Removed ID, target mapping, wrong records, fill NaN

49998 x 18

model_cleaned.csv

49998 x 4 → df_numerical

49998 x 4 → df_categorical

49998 x 10 → df_review

---

df_numerical

49998 x 4

Log transformation ◇

49998 x 6

df_numerical

49998 x 6

Correlation test ◇

49998 x 3

df_numerical

49998 x 3

---

df_categorical

49998 x 4

◇ Chi Square test

49998 x 2

df_categorical

49998 x 2

◇ One hot encoding (drop_first)

49998 x 2

dummies

49998 x 2

---

df_review

49998 x 10

◇ Chi Square test

49998 x 10

df_review

49998 x 10

---

selected_features.csv

49998 x 15

StandardScaler() ◇

49998 x 15

final.csv

49998 x 15

◇ DownSampling of Zeros

41004 x 15

final_down.csv

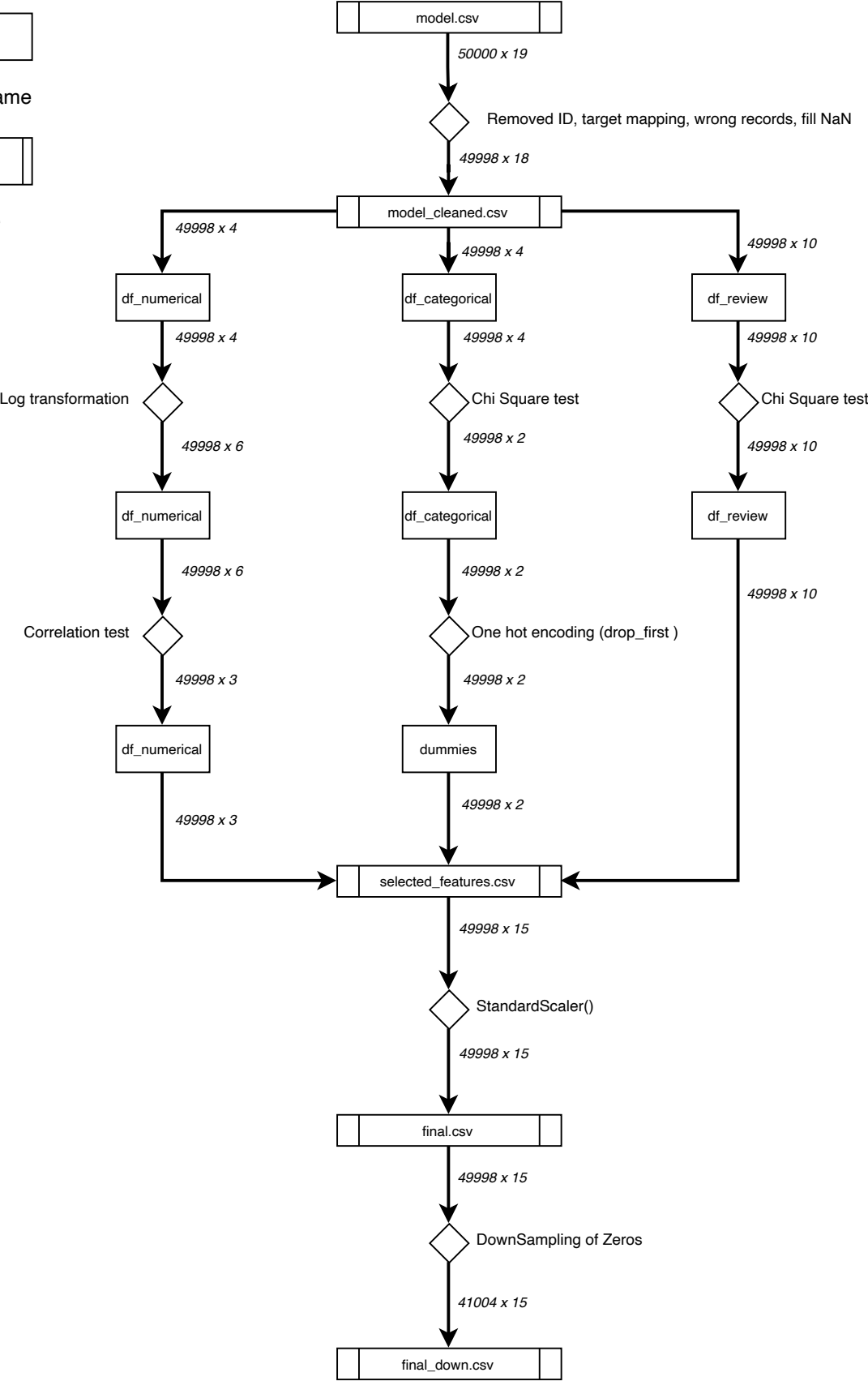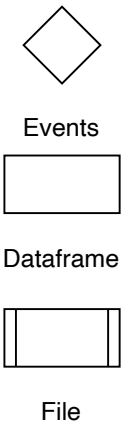
Note: in notebooks there are duplicate of target variable for data viz purposes. In this schema are reported NO duplicated.