Índice del Proyecto de Ciencia de Datos - Data Cleaning (Jupyter Notebook 1)

1. Introducción

1.1. Objetivos del notebook

1.2. exportación de librarias utilizadas

2. Carga y procesamiento del dataframe principal (app\_train)

2.1. Carga de app\_train

2.2. Eliminación de variables con 30% o más de valores nulos

2.3. Eliminación de variables irrelevantes

2.4. Conversión de variables categóricas a binarias

2.5. Eliminación de variables object faltantes

2.6. Eliminación de variables con alto porcentaje del mismo valor

2.7. Identificación y transformación de outliers a NA

2.8. Relleno de NAs con el método ANOVA

2.9. Guardado del nuevo CSV

3. Carga y procesamiento de dataframes secundarios

3.1. Bureau y bureau\_balance

3.1.1. Carga de Bureau

3.1.2. Eliminación de variables con 30% o más de valores nulos

3.1.3. Eliminación de variables irrelevantes

3.1.4. Conversión de variables categóricas a binarias

3.1.5. Eliminación de variables object faltantes

3.1.6. Eliminación de variables con alto porcentaje del mismo valor

3.1.7. Identificación y transformación de outliers a NA

3.1.8. Relleno de NAs con el método ANOVA

3.1.9. Creación de variables dummies

3.1.10. Carga de bureau\_balance

3.1.11. Fusión de los dos CSVs

3.2. previous\_application

3.2.1. Carga de previous\_application

3.2.2. Eliminación de variables con 30% o más de valores nulos

3.2.3. Identificación de variables con constantes o quasiconstantes

3.2.4. Eliminación de columnas no relevantes

3.2.5. Transformación de variables seleccionadas a dummies

3.2.6. Eliminación de variables quasiconstantes

3.2.7. Reemplazo de outliers con nans

3.2.8. Limpieza ANOVA

3.2.9. Guardado del nuevo CSV

3.3. POS\_CASH\_BALANCE

3.3.1. Carga de pos\_cash

3.3.2. Eliminación de variables con 30% o más de valores nulos

3.3.3. Identificación de variables con constantes o quasiconstantes

3.3.4. Eliminación de columnas no relevantes

3.3.5. Transformación de variables seleccionadas a dummies

3.3.6. Eliminación de variables quasiconstantes

3.3.7. Graficar variables para exploración visual

3.3.8. Reemplazo de outliers con nans

3.3.9. Limpieza ANOVA

3.3.10. Guardado del nuevo CSV

3.4. credit\_card\_balance

3.4.1. Carga de credit\_card

3.4.2. Eliminación de variables con 30% o más de valores nulos

3.4.3. Identificación de variables con constantes o quasiconstantes

3.4.4. Eliminación de columnas no relevantes

3.4.5. Transformación de variables seleccionadas a dummies

3.4.6. Eliminación de variables quasiconstantes

3.4.7. Graficar variables para exploración visual

3.4.8. Reemplazo de outliers con nans

3.4.9. Limpieza ANOVA

3.4.10.Guardado del nuevo CSV

3.5. installment\_payments

3.5.1. Carga de installment\_payments

3.5.2. Eliminación de variables con 30% o más de valores nulos

3.5.3. Identificación de variables con constantes o quasiconstantes

3.5.4. Eliminación de columnas no relevantes

3.5.5. Transformación de variables seleccionadas a dummies

3.5.6. Eliminación de variables quasiconstantes

3.5.7. Graficar variables para exploración visual

3.5.8. Reemplazo de outliers con nans

3.5.9. Limpieza ANOVA

3.5.10. No se guarda este último CSV

Data Preprocessing (Jupyter Notebook 2)

Introducción

* 1. Objetivos del preprocesamiento
  2. Exportación de librerias utilizadas

Preprocesamiento del dataframe bC (Bureau y Bureau Balance)

2.2. Carga de bC

2.3. Graficado de histogramas

2.4. Eliminación de columnas constantes y quasiconstantes

\*TODO\*2.5. Creación de variables dummies a partir de variables numéricas positivas, 0 y negativas

2.6. Eliminación de variables correlacionadas

2.7. GroupBy

2.8. Reset index

2.9. Guardado del nuevo CSV

Preprocesamiento del dataframe train\_app

3.1. Carga de train\_app

3.2. Graficado de histogramas

3.3. Eliminación de columnas constantes

3.4. Eliminación de columnas cuasiconstantes

3.5. Guardado del nuevo CSV

Preprocesamiento del dataframe installment\_payment

4.1. Carga de installment\_payment

4.2. Histograma

4.3. GroupBy

4.4. Guardado del nuevo CSV

Preprocesamiento del dataframe prev\_app

5.1. Carga del CSV prev\_app

5.2. Eliminación de variables correlacionadas

5.3. Eliminación de columnas constantes

5.4. Graficado de histogramas

5.5. GroupBy

5.6. Guardado del nuevo CSV

Preprocesamiento del dataframe creditCardBalance

6.1. Carga del CSV creditCardBalance

6.2. Histogramas

6.3. GroupBy

6.4. Guardado del nuevo CSV

Fusión de todos los dataframes

7.1. Carga de todos los CSVs

7.2. Merge entre todos los dataframes

7.3. Rellenar 0 los valores nulos “TODO es correcto rellenar con ceros Warning”

7.3. Guardado del CSV final

Machine Learning Process (Jupyter Notebook 3)

Introducción

1.1. Propósito del proceso de Machine Learning

Preparación de los datos

2.1. Carga de librerias necesarias

2.2. Carga del dataframe

2.3. Eliminación de las variables KEY

Balanceo de datos

3.1. Balanceo de los datos con submuestreo

Separación de variables

4.1. Separación de X y y en diferentes variables

División de datos

5.1. División de datos en conjunto de entrenamiento y prueba

Creación y evaluación de modelos

6.1. RandomForestClassifier

6.2. Comparación de diferentes modelos

6.2.1. GradientBoostingClassifier

6.2.2. XGBoost

Conclusiones