**EDA – Categorical Features**

1. **General updates**

* Dropped features in **train\_df** that is not in **test\_df (except the target “Claim Injury Type”)**
* Dropped 3.28% of the training data that has NAs on the target

1. **Features after item 1**

|  |  |
| --- | --- |
| *Alternative Dispute Resolution* | * Not isna() * Dropped lines == “U” (annomaly) * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *Attorney Representative* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *Carrier Name* | * Not isna() * MultiEncoder(frequency) |
| *Carrier Type* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (int) |
| *Claim Injury Type* | * Isna() treated in item 1 * Target feature * MultiEncoder(Binary) – LabelEncoder (int) |
| *County of Injury* | * Not isna() * MultiEncoder(frequency) |
| *COVID-19 Indicator* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *District Name* | * Not isna() * MultiEncoder(frequency) |
| *Gender* | * Not isna() * Replace(“U”, “X”) » domain: (“M”, “F”, ”X”) * MultiEncoder(frequency) |
| *Industry Code Description* | * Strongly correlated with “*Industry Code*” * Description feature * Dropped |
| *Medical Fee Region* | * Not isna() * MultiEncoder(frequency) |
| *OIICS Nature of Injury Description* | * Only NaN * Dropped |
| *WCIO Cause of Injury Description'* | * Strongly correlated with “*WCIO Cause of Injury Code*” * Description feature * Dropped |
| *WCIO Nature of Injury Description'* | * Strongly correlated with “WCIO Nature of Injury Code” * Description feature * Dropped |
| *WCIO Part Of Body Description'* | * Strongly correlated with *“WCIO Part Of Body Code”* * Description feature * Dropped |

1. **Categorical and also Numeric Features**

|  |  |
| --- | --- |
| *Industry Code* | * 1,73% NaN |
| *WCIO Cause of Injury Code* | * Not isna() |
| *WCIO Part Of Body Code* | * Not isna() |
| *WCIO Nature of Injury Description* | * Not isna() |
| *Zip Code* | * not.isnumeric() to NaN * Not isna() (fillna(0)) |

Obs.: Missed Values on *Industry Code* filled after split (train & val) for train data integrity maintenance (Rui).