Report for Advanced Analytics in Business

Cheung Wai Chun, r0817438

David Badajkov, r0604517

Ana Maria Giraldo Vargas, r0822450

Sonia Rocio Socadagui Casas, r0823960

Marcela Lopez Viveros, r0773141

April 29, 2021

Contents

Assignment 1

Feature engineering

The (training) dataset for Assignment 1 consists of 55463 observations and 78 features. The number of features. It is easy to observe that there are a lot of missing values and categorical or date features in the dataset. In this section, we would discuss the strategies used in handling such problems.

Missing values

As shown in figure 1, there are 55 feature which contain missing values, and 24 out of them contain more than 80% of missing values. For those features with a high proportion of missing values (more than 50%), missing values are treated as an extra category. By treating missing values as an extra category, the information of the non-missing entries of those features can be retained and learnt by the model, whereas removal of those features may lead to a loss in information or pattern. For those features with a lower proportion of missing values, imputation techniques can be applied to estimate their possible values.

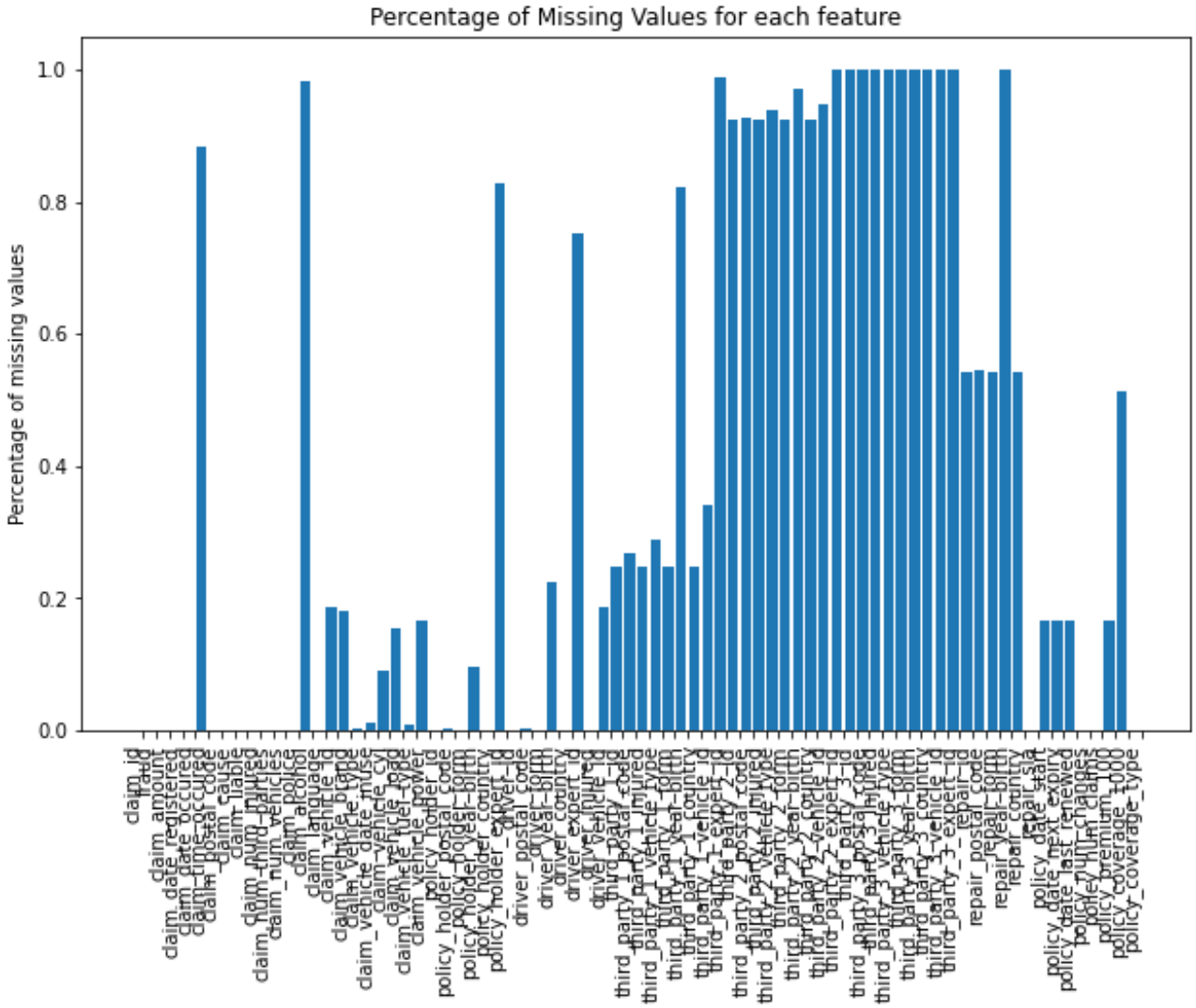

Figure 1: Proportions of missing values for each feature

Date features

Some features are available in form of date. In the dataset, the following features are in terms of date: `claim_date_registered`, `claim_date_occured`, `claim_vehicle_date_inuse`, `policy_date_start`, `policy_date_next_expiry` and `policy_date_last_renewed`. However, date itself is not suitable to serve as an input for some machine learning models. Hence, as summarized in table 1, we construct another set of features which are more interpretable and meaningful based on these date features.

Apart from the aforementioned features, features related to birth year are also considered as date features. They are `policy_holder_year_birth`, `driver_year_birth`, `third_party_1_year_birth`, `third_party_2_year_birth`, `third_party_3_year_birth` and `repair_year_birth` in the dataset. By subtracting them from the year of `claim_date_occured`, we obtain age-related features.

Constructed features	Descriptions
`days_before_registered`	The number of days between `claim_date_registered` and `claim_date_occured`.
`days_before_occured`	The number of days between `claim_date_occured` and `claim_vehicle_date_inuse`.
`policy_length`	The number of days between `policy_date_last_renewed` and `policy_date_start`.
`policy_claim_length`	The number of days between `policy_date_next_expiry` and `claim_date_occured`.

Table 1: Featurization of date features

Data quality issue

During data cleaning process, we discovered some problems with regards to data quality. For example, the instance with claim id 62780 has an invalid value for the year of `claim_vehicle_date_inuse`. Such observations may be due to input error in manual data entering process. However, they can be hard to observe in general.

```
[15]: # Input error in year of claim_vehicle_date_inuse
      data[data['claim_vehicle_date_inuse'] > 202012]
```

[15]:

n_vehicle_id	claim_vehicle_brand	claim_vehicle_type	claim_vehicle_date_inuse	claim_ve
J5ZWJmMjQ	CITROEN	car	270505.0	

Figure 2: Invalid value of `claim_vehicle_date_inuse`

Data binning

Binning continuous features can help incorporating missing values and extreme values in a more natural way as they can be reformulated as categorical features. It is useful for those continuous features with a high proportion of missing values or highly right-skewed. For age-related features, they are binned by age groups with equal intervals, referencing to their histograms.

Exploratory data analysis

Model building

Interpretation

Reflection

Assignment 2

Assignment 3

Assignment 4