uday sai

4 Followers About

Feature Engineering: Dealing with Outliers and Rare Labels

uday sai Apr 19, 2020 · 6 min read

In this article we will be going through some of the techniques to deal with the outliers of a numerical variable and rare labels of a categorical variable. Please refer the below link to read my article on “EDA :Detecting Anomalies”

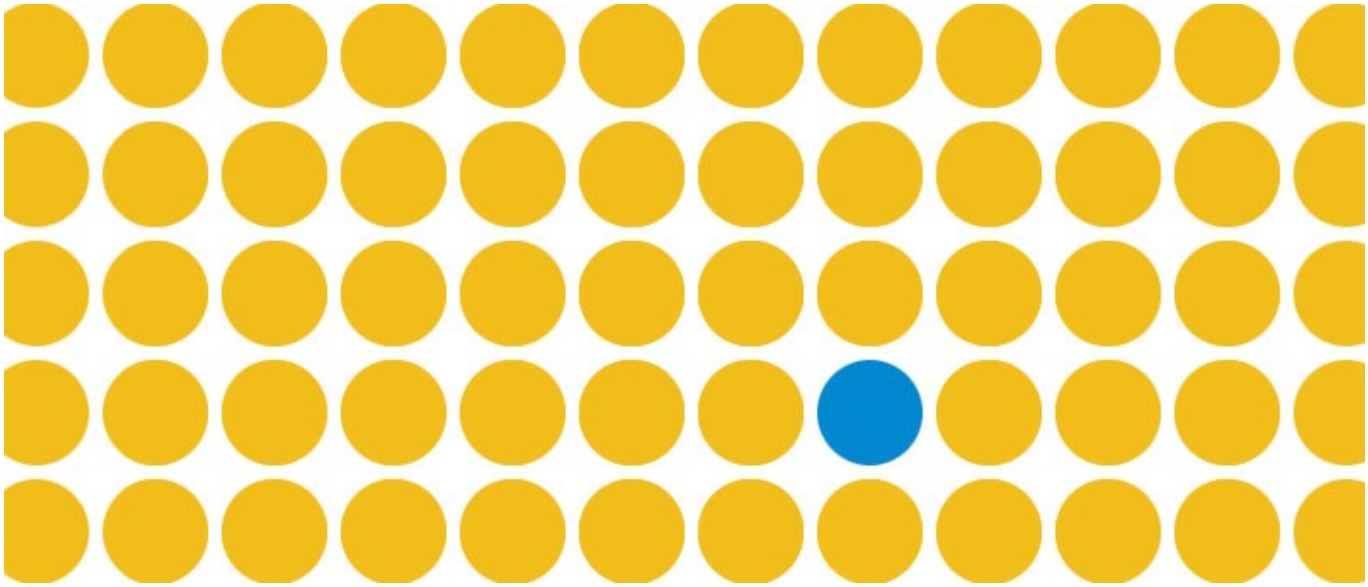

Image Source: https://blog.sonlight.com/how-homeschooling-affects-conformity.html

How to perform EDA which complements our Data Science story: Detecting Anomalies

“Data Anomalies are not always noise in our data, they might be inbuilt

Let us look in to some of the techniques to deal with the outliers. In using these techniques it is very important to remember that these techniques should be used only in case of true outliers and not in case of genuine outliers. For example, in a dataset which contains salary column there will be few values of high salary. In these cases the columns should be rightly processed using any of the feature scaling techniques.

> *Deletion of Outliers*
>
> *Mean/Median Imputation*
>
> *Top-Coding, Bottom-Coding and Zero-Coding*
>
> *Discretisation*

Deletion of Outliers

In this technique we will delete the rows which contains outliers after we make sure that the outliers are potential errors due to data collection process or any other methodologies used in the subsequent steps. The deletion must be only done after confirming with the domain experts/product owners that the values are true outliers. Some times outliers are used to mask the null/missing values. In this case the outliers must be treated as missing values and processed accordingly.

Mean/Median Imputation

In this technique we will be replacing the outliers with mean/median imputation. This method is equivalent to simple mean/median imputation of missing values. In the previous step we already discussed that some times outliers are used to mask the missing/null values. In this case the outliers must be treated as missing values and processed accordingly.

Top-Coding, Bottom-Coding and Zero-Coding

This statistical technique is widely used in econometrics in which we cap the maximum of a distribution at an arbitrarily set value. The masking of the values which are greater

Top-coding is common practice in survey data, before it is released to the public. It is used to preserve the anonymity of respondents. For example, high earners may be easily identifiable by their earnings. Thus, by implementing top-coding, that outlier is capped at a certain maximum value and therefore looks like many other observations, it is not uniquely identifiable any more. Top-coding can be also applied to prevent possibly-erroneous outliers from being published.

Bottom-Coding

This technique is analogous to top-coding on the left side of the distribution. In this method all the values below a certain threshold are capped at threshold. This technique is commonly used in the surveys data cleaning. Top-coding and bottom-coding are indeed used in practice to remove outliers of variables and therefore prevent model over-fitting.

Zero-Coding: It is a special case in the bottom-coding in which we set the threshold as zero. For example, if the variable “age” has negative values, we will set the threshold value as zero. Any observation with a value under zero must have been introduced by mistake.

Note

Top-coding may affect estimates of the standard errors of the variable, or change the variable distribution, by censoring those values at the far end of the tails.

Discretisation

In this technique we will transform a continuous variable into a discrete variable. For example, let us consider if we have a variable “Age” which takes many continuous values, we can categorize the values under bins 0–20, 21–40, 41–60, 61–80, >80. By this process we will convert the outliers in to one category. This technique will be used top prevent over-fitting.

Dealing with Rare Labels

to deem a label as a rare label must be done after thorough inspection in EDA, in terms of response variable variation. In implementing this technique we may encounter three scenarios usually:

> *One predominant Category*
>
> *Small number of Labels*
>
> *Categories with high cardinality*

One Predominant Category Label

In this scenario we encounter the categorical variable with one dominant label and other labels of small proportion. The best example is the 'Street' variable in the advanced house price prediction dataset (downloaded from kaggle).

```
print(df['Street'].value_counts()/len(df))
```

```
Pave    0.99589
Grvl    0.00411
Name: Street, dtype: float64
```

Image: Checking for labels of 'Street' variable

In this technique we will initially check for the response variable patterns against the labels of the categorical variable. If the other variables does not affect the response variable significantly the other labels which appear in small proportion are replaced with the dominant label. This prevents the over-fitting of the model and accounts for generalization.

Small number of Labels

In this technique we will usually convert the categories which are present less than 10 percent to rare categories. However we must choose the threshold after a thorough inspection in EDA. The rare categories must be grouped only if they have same effect on the response variable. If the response is erratic then the labels must be grouped based on the response pattern. This technique proves to be instrumental in the model generalization.

High Cardinality Categorical Variables

In this scenario we have categorical variables which has high cardinality meaning large number of labels and this is the case where our technique of “rare” value imputation comes to rescue. However as discussed before we need to substitute the rare labels only after carefully checking the patterns with the target variable.

In this context if the categorical variable has high level of cardinality like 30–40 in which 25–35 of the labels appears in the range of 2–8 percent then this technique demonstrates it’s core purpose. For example, we can consider the mercedes benz dataset downloaded from kaggle.

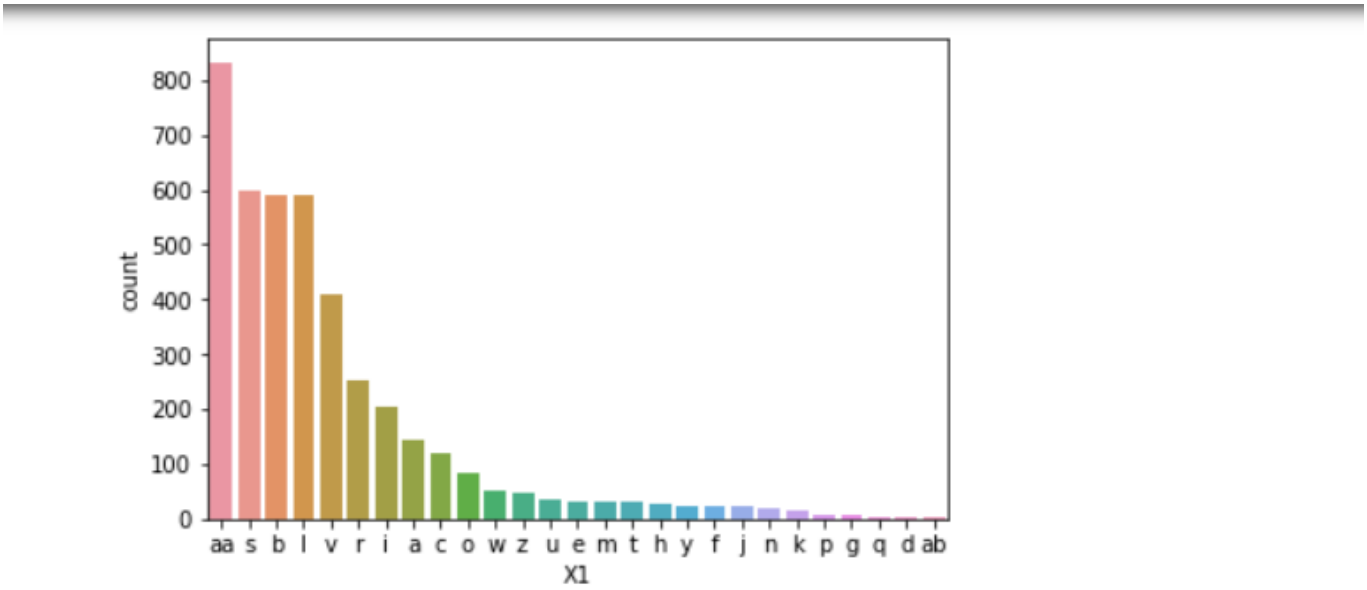

Image: Categorical Variable with high cardinality

In the above case if we can replace the labels which occurs very minimal number of times we can categorize these labels as ‘rare’ label. By doing so we are minimizing the noise and preventing the model to overfit.

Conclusion

In the above techniques we had seen how to deal with outliers in case of numerical variables and rare labels in terms of the categorical variables. The important point to note here is to note how these outliers and rare values impact the target variable. The

On the other hand, the categorical variables with rare labels must be processed as according to any of the above mentioned context based on the variation with the response variable.

Please refer my github profile to check on the entire code.

References

udaysai50/Full-Stack-Data-Science

Permalink Dismiss GitHub is home to over 40 million developers working together to host and review code, manage...

github.com

krishnaik06 - Overview

Dismiss Sign up for your own profile on GitHub, the best place to host code, manage projects, and build software...

github.com

How Homeschooling Affects Conformity | Sonlight Homeschooling Blog

Henry Cate asked: Is group pressure what motivates us to root for our home team? He has an excellent (and funny) video...

blog.sonlight.com