uday sai

4 Followers About

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

uday sai Apr 14, 2020 · 4 min read ★

> "Data Anomalies are not always noise in our data, they might be inbuilt patterns which needs critical examination and often accounts for comprehending data in a better way"

In this article I will be implementing the EDA in python to detect anomalies. The anomalies encompasses missing values, the outliers of a numerical variable and rare labels of a categorical variable. In a real-world data science project we often encounter explanatory variables which has null values/ NAN's, variables with high cardinality. In this case intuitively we may feel that these contexts does not contribute for information and in turn adds noise. However, we cannot arrive at a conclusion that outliers, rare labels always account for noise. In fact its an opportunity to understand data by performing robust analysis.

Missing Values

Outliers in a Numerical Variable

Let us explore one at a time.

Missing Values:

The first task in detecting anomalies is to check for missing values, which might occur in any of the explanatory variables. After identifying missing values we need to question the occurrence of missing values and categorize them. There are three common reasons for occurrence of missing values like Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

MCAR: This case occurs when the occurrence of missing values is equally likely for all the instances of data i.e., if a feature is missing for a instance it is completely by chance and missing value is not due to any other explanatory feature or target variable.

MAR: This context occurs when the probability of missing values is dependent on other features. For example, in a survey women are tend not to disclose their weight when compared to men. If this the data that we are analyzing then the missing values in weight column can be imputed using other variables such as height, age etc…

MNAR: This case occurs when the reason for missing values is firmly known. For example, in a telecom customer churn data the column 'totalcharges' might contain missing values. This is because the 'tenure' column associated with same datapoint is zero. In this cases utmost care should be taken to impute these missing values.

Outliers in a Numerical Variable

The outliers of a numerical variable can be detected in two ways:

a.) if a variable follows a perfect Gaussian distribution then outliers can be identified as points which are out of (3*sd-mean,3*sd+mean) range (where sd is standard deviation of the feature)

b.) The other technique is to find the inter-quartile range (IQR) and identifying the data points outside range (Q1–1.5*IQR,Q3+1.5*IQR) as outliers and points out of range (Q1–3*IQR,Q3+3*IQR) as strong-outliers.

example of Titanic dataset (borrowed from kaggle) and check for outliers in “Fare” column. We can see many outliers in the below picture in the boxplot.

```
In [18]: sns.boxplot(df['Fare'],orient="v")
         plt.xlabel("Fare")
         plt.title("Outlier Detection")
         plt.show()
```

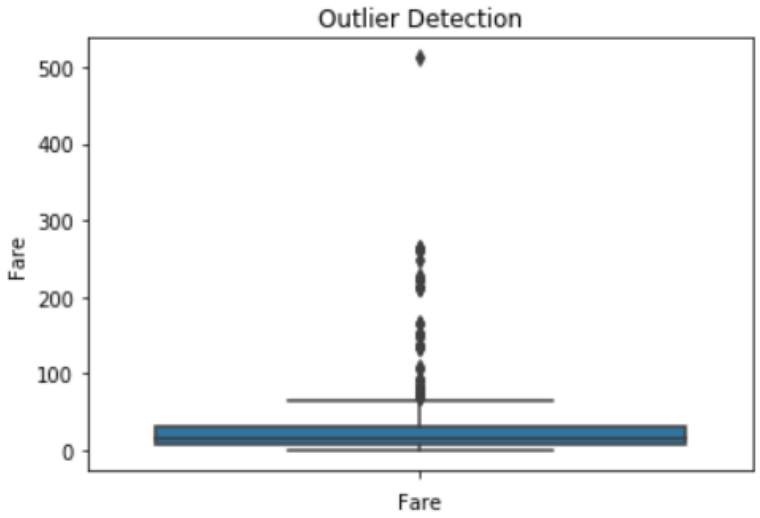

Image: Outlier Detection

But if we carefully go through the rows which contain these outliers, we can make out that these outliers are because of combined fare prices. In the below picture we can see multiple name and cabin values for rows containing outliers.

Image: Outlier Identification

Rare Labels in a Categorical Variable

Often times, in a real-world dataset we come across categorical features which contain rare labels meaning the labels are associated with fewer data points in our dataset. In this scenario we need to check thoroughly if these labels really add value to our target variable. Let us look at the below example where we consider mercedesbenz dataset from kaggle, the variables are intentionally masked by the company. Here we can see that variables 'X1', 'X2', 'X3', 'X6' contain multiple labels and fewer instances associated with some labels.

Image: Rare Labels

In this cases we can try to group the rare labels if the associating pattern of response variable remains same for these group. However, we cannot conclude without consulting the domain experts.

Conclusion

These are some of the mechanisms to identify anomalies and finding out if they needs special attention or simple deletion.

References

udaysai50/Full-Stack-Data-Science

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

krishnaik06 - Overview

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

github.com

Eda Python Machine Learning Data Science Data Visualization