A Study of the Mushroom Dataset

Léo Jacqmin
May 11, 2020

1 INTRODUCTION

This is the classical problem of predicting a binary outcome: is a mushroom poisonous or not? Using the Mushroom Dataset [1], we survey the various steps of a data science project to find out which features are most indicative of a poisonous mushroom and see which machine learning models perform best for this task.

2 DATASET

This dataset is known as the mushroom dataset. It includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. There are 22 attributes, which are all categorical, and 8124 training instances. The goal is to predict the binary class 'poisonous' or 'edible'. While this dataset was originally curated more than 30 years ago by the UCI Machine Learning repository, mushroom hunting is still relevant to this day and the data provides for a good case-study of modeling with categorical variables only.

3 DATA PREPARATION

After collecting the data, the first step is to prepare it for further analysis. The cleaning process in this project includes:

[1] `https://archive.ics.uci.edu/ml/datasets/Mushroom`

- Deleting columns that contain a single value: They add no information for this classification task.
- Checking for missing values: Any missing value has to be dealt with. In this case, there are none.
- Converting categorical values to ordinal values using a one-hot encoding: Each categorical feature with n-categories possible values is transformed into n-categories binary features, with one of them 1, and all others 0. This prevents some estimators from interpreting categories as being ordered.

4 Exploratory Data Analysis

This process can be likened to getting to know the data. By analysing how the features relate to each other, we gain crucial knowledge about the data which helps training efficient models.

One important thing to consider first is whether the dataset is balanced, i.e. whether the dataset contains the same number of samples from the positive and negative class. In this case, the dataset is balanced.

Next, Cramer's V statistical test allows us to measure the strength of the association between categorical variables. It is based on Pearson's chi-squared statistic and has the advantage of giving good norming from 0 to 1 regardless of table size. Correlated features in general don't improve models, but they affect specific models in different ways and to varying extents:

- Solutions from linear models may vary greatly due to multicolinearity.
- Highly correlated features can mask the interactions between different features which are detected by random forests.

More generally, a simpler model is preferable, and, in some sense, a model with fewer features is simpler, which leads us to the next section.

5 Feature Selection

Feature selection is the automatic selection of features that are most relevant to a predictive modeling problem. The objective of feature selection is three-fold:

- improving the prediction performance of the models
- providing faster and more cost-effective models
- providing a better understanding of the underlying process that generated the data

Using filter methods such as the Chi-squared test and information gain and selecting the top 25 features results in a worsened accuracy. This can be explained by the fact that the number of features in the dataset is relatively low to begin with.

Another filter method is to ignore features with low variance. These features can be dismissed as they do not contribute to the model's accuracy. For the same reason stated above, we keep all features.

6 Data Modeling

When it comes to machine learning algorithms, there is no one-size-fits-all. Depending on the task and the data, different models might perform better. Thus, the best approach is to try various models, tune them and compare how they perform.

It turns out this problem is easily solved by modern classifiers such as SVC and XGBoost as they all achieve 100% accuracy. We observe that different models use the data in different ways, though there are some similarities in what features are considered most important.

We then study how a Decision Tree Classifier makes predictions. This model has the advantage of being highly interpretable and is well-suited for categorical variables. In this tree, each node represents a feature and the leaves correspond to the predicted outcome. Interestingly, only three features account for about 90% of the information the Decision Tree Classifier needs to achieve 100% accuracy.

7 Hyperparameter Tuning

Although the prediction are already completely accurate, we perform hyperparameter tuning for the sake of demonstration. This final step helps bring up the performance of predictors. We can find which combination of hyperparameters performs best by using a randomized search which samples from a set of pre-defined hyperparameters and fits various models with different combinations of paramaters. The most accurate model can then be retrieved to be used.

8 Conclusion

In this study of the Mushroom Dataset, we have surveyed how to approach a classification problem with categorical features. While this dataset is rather old and modern predictors easily reach 100% accuracy, it still serves as an interesting basis for exploratory analysis.