Capstone Proposal

Machine Learning Engineer Nanodegree

Title: Evaluation of Nursery School Applications

Alphonso Jo Stephen
February 7th, 2019

Proposal

Domain Background:

Nursery school, pre-primary school, playschool or kindergarten, is an educational establishment or learning space offering early childhood education to children before they begin compulsory education at primary school. It may be publicly or privately operated, and may be subsidized from public funds.

In an age when school was restricted to children who had already learned to read and write at home, there were many attempts to make school accessible to orphans or to the children of women who worked in factories.

With development of science and technology and with improved civilization, the need for basic education has increased which in turn lead to high number of schools and colleges. But here we are only discussing about Nursery school which is the first door towards education

Here I am finding out whether a child is accepted into a nursery school or not by evaluating the accuracy and acceptance

Database reference: https://archive.ics.uci.edu/ml/machine-learning-databases/nursery/nursery.names

Problem Statement:

The main objective of this project is to evaluate all the applications that are enrolled into a nursery school and identify the children who are capable of getting an admission based on some factors. Classification technique like Decision Tree classifier will be implemented and compared upon standard metric like accuracy. Pipeline is also used to avoid any data leakage.

There are several factors which are based upon to accept the application form like parents, social status, health checkup, housing status, financial stability. Nursery School prediction predicts the acceptance of children into the kindergarten.

Datasets and Inputs:

Dataset Link: https://archive.ics.uci.edu/ml/datasets/Nursery'

In this project I have used 9 attributes and around 1200 trained and test data to evaluate the accuracy score. And this accuracy of classifiers is found out by 5-fold cross validation.

Data Set Information:

Nursery Database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It was used during several years in 1980's when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation. The final decision depended on three sub problems: occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The model was developed within expert system shell for decision making DEX (M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.).

Features Information:

- parents - usual or pretentious (real or adapted parents)
- has_nurs – has proper nursing or not
- form – application form complete or incomplete
- children – number of children
- housing – status of house (convenient or inconvenient)
- finance – financial stability(convenient or not)
- social - problematic or not
- health – health status
- target – final decision (recommended or not)

Solution Statement:

In this project, I am trying to predict the eligibility criteria of children who applied for this nursery school. This can be achieved by evaluating all the features mentioned above and finding accuracy using classification techniques like Decision Tree Classifier, Random Forests, and Logistic Regression. I will explore the data set with matplotlib.py, seaborn libraries to plot. Visualization helps to understand the model more clearly.

Benchmark Model:

Here we compare the final model with the remaining models to see if it got better or same or worse. The accuracy is compared among the models and the best model is selected. I think logistic regression model can be set as the benchmark model and I'm sure that the final solution would outperform the Benchmark model

Evaluation Metrics:

Accuracy:

It is the number of correct predictions made by the model over all kinds of predictions made

$$\text{Accuracy} = (TP+TN) / (TP+FP+FN+TN)$$

Accuracy is a good measure when the target variable classes in the data are nearly balanced. Here I am predicting the accuracy score for selected models and the best score model is selected.

Project Design:

- **Preprocessing:** It is the first step to read the dataset and clean the data i.e. removing unwanted data or identifying null values. If any null values exist, we replace them with constant values or removing duplicates.
- **Exploration:** Visualizing the dataset, detect outliers, replacing a missing value and cleaning the dataset, splitting training dataset into training and testing sets and checks for any correlation among the features using heatmap.
- **Prediction:** Here we predict the eligibility criteria of children who applied for nursery school by finding accuracy using different classification models

Finally, I declare that the model with the highest accuracy score on both training and testing datasets will be concluded as the best model for evaluating the nursery school application form.