Machine Learning Engineer Nanodegree

Capstone Proposal

Perihane Youssef Yasser

February 2, 2019

Domain Background

Day nurseries are a great option if you want your child's learning and care to have structure. Staff are trained to create a safe, happy and stimulating environment where your child can play and develop. Your child will be able to do a wide variety of activities at day nursery. These activities will be designed to encourage your child's social, creative, communication and listening skills and his physical development[1].

The problem lies in choosing the best nursery for your child, where he can make use of his potentials and skills and feel accepted among his peers. That's why nurseries put such great effort in interviewing new children and set very strict criteria on evaluating and accepting them depending on the child and his parents financial and social standings and what they seek in a "day care" facility for their child.

Figure 1: infants and toddlers are spending time in a nursery[2]

Problem Statement

Our main concern is to decide if a child's application in the nursery will be accepted or rejected. Since that any poor choice will have a negative effect on the child and his colleagues as children are so sensitive in their first years, during which learning depends mainly on observing and imitating.

So our ranking system will evaluate every child separately based on some attributes as: The social picture and health of the child's family, the financial status of his parents, whether the child is supported by one or two parents, etc.

Objective:

To classify the child as recommend, priority, not recommend or very recommend using a machine learning model that can predict the category of each child based on his/her features.

Datasets and Inputs

The dataset used for this project is "Nursery Data Set", I found it on UCI Machine Learning Repository[3]. 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.

It was published in 1997 and it has 12960 rows and 8 columns, each column represent one feature of a child and his parents (a piece of information). Each row represents a data point: one child observation. The target variable to be predicted in this problem is the "NURSERY" variable. It contains four values: recommend, priority, not_recom (not recommend) and very_recom (very recommend). The rest of columns represent information about occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. All the columns contains categorical variables and they will be encoded to numerical values to be able to feed them to different machine learning models, and there are no missing values.

Solution Statement

This project will depend on developing different machine learning models and choosing the one with the most accurate classification of the child's application in the nursery. The model with best scores during training and testing will be used for further prediction of the status of each child.

Benchmark Model

The benchmark model is a simple "Logistic Regression" classifier. It is simple and doesn't require any hyper-parameters to specify beforehand.
Logistic regression is by far the most widely used classifier in real-world applications, so it would be interesting to outperform it and beat it's score.

Evaluation Metrics

The main performance measure that will be used is testing accuracy, testing accuracy will ensure that future observations will be classified correctly. Other performance metrics will be used as needed.

Project Design

I will follow some steps to find the best model and accurately predict each data point class.

1. Exploratory data analysis
   Understand and making sense of the data first and try to gather as many insights from it. Performing initial investigations to discover patterns, to know the number of categories in each columns and to apply some statistics measures.

2. Data preprocessing
   Based on the exploration of data, I will transform raw data into an understandable format, convert categorical values to numerical ones and decide if any further enhancements are needed.

3. Split the data set
   Split the data into training and testing sets.

4. Build benchmark model
   Build the logistic regression model and evaluating it.

5. Build Models
   Build different machine learning models, training these models using the training set. (Note: Model types will be chosen during implementation)

6. Evaluation and Benchmarking
   Use the testing set to evaluate each model by different performance metrics, then choose the best model and compare it by the benchmark model.

7. Validation:
   Hyper-parameter tuning and validation for the best model.

Note: I may add some steps if it's necessary during project building.

References

[1] "Pros and cons of day nurseries," https://www.babycentre.co.uk/a537552/pros-and-cons-of-day-nurseries

[2] "How to choose quality child care," https://www.zerotothree.org/resources/84-how-to-choose-quality-child-care

[3] UCI Machine learning repository: "Nursery Data Set," https://archive.ics.uci.edu/ml/datasets/nursery?fbclid=IwAR0qzA8oKVP-4Sv9WRggMSwCUr8nqtj7AnJRMGGM4i5cLQnBUUg3loZxOpQ