Task System Analysis Documentation

The various steps involved in the completion of the project are:

1. Business Understanding

In this, the business problem and the motivation of that problem had to be understood in order to act accordingly further in the project.

While employers get high response to their posting, it is difficult to go through a high number of applications for the employers. They might need to go through high number of applications to shortlist the most relevant candidates. Hence an intelligent matching algorithm can help our users to get better experience and enhance the chances of meaningful profile matches.

This business problem was provided. Then further research on the particular domain was done using different resources over the internet.

2. Data Understanding

In this step, the roles and meanings of different variables present in the Dataset had to be understood.

Internship.csv - includes the details of all the internships posted on Internshala. These details are filled by the company floating the Internship. Each row represents one internship. It has 286 columns and 6899 rows. This includes various attributes about the internships like location, duration, start_date of internship etc.

Student.csv - includes details of the students applying for the internship. These details have been filled by the student. Each row represents an experience of the student. In case the student has not filled any experience, there would be only one row containing details of student. It has 19 columns and 151191 rows. These include type_of_institute, current_year, academic performance of the student etc.

test.csv & train.csv - include the application details (as applied by student) and the shortlist outcome

*Any student is free to apply for any internship on the portal.

3. Data Preparation

In this step data cleaning is performed in order to prepare data for further analysis and for building the machine learning model. The data which is obtained may need to be processed before it can be actually used, like there may be some values missing which need to be filled otherwise they will cause problem when doing the analyses on the data.

4. Modeling

In this step the appropriate machine learning model is selected, then dataset is divided into the training and test sets, and then finally the model is trained on the training data in order to get the best results.

Since our problem is a classification problem and has a very large size of more than 100k samples , we decided to use the technique of Gradient descent for which the most suitable algorithm is Gradient boosting which is able to solve the problem of large size, handling data of mixed type and missing values, robust to outliers in input space and has a good interpretability and predictive power.

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

5. Evaluation

In this step, the model is evaluated against the test set using various performance metrics like looking at R-squared, Adjusted R-squared, Residual Plots, etc.

We obtained a model report with accuracy of 87.58 percent, which is the best fit.

6. Deployment

The model is then finalized after getting satisfied results in the previous step and finally delivered and presented along with documentation.