Responsible Data Science Course Project

Claire Saint-Donat, Xiangyue Wang

Center for Data Science, New York University

Spring 2022

Contents

1 Introduction

The purpose of this project is to build an interpretability tool for an Automated Decision System (ADS) which we have chosen to audit.

Automated Decision Systems are in widespread use in government and industry, and a number of efforts are currently underway to regulate them. New York City recently passed a law (Local Law 49 of 2018) that compels the development of procedures and recommendations that City agencies should follow when explaining the operation of an ADS to the public, and demonstrating that an ADS does not discriminate against individuals based on membership in protected groups.

In this project, we attempt to help NYC and other municipalities by designing a“nutritional label” (similar to a label used to evaluate food products) for an algorithmic system of our choice.

2 Background

In recent years, the growing demands for jobs in data science have driven the creation of a myriad of degree programs, massive online open courses, online certification programs, and boot camps. While the abundance of those learning resources has made data science accessible to a wide range of people looking to enter the field, the variety of credentials has also made it difficult for companies to evaluate the qualifications of job seekers. The process of hiring data scientists is often both expensive and time-consuming. Therefore, companies have incentives to hire qualified data scientists and keep them from leaving.

As as a result of this trend, the Automated Decision System (ADS) that we chose to examine is from the Human Resources and hiring space. In particular, we wish to examine a top entry from an HR Analytics Kaggle competition published here. The source of these data come from a company active in the big data and data science industries that is looking to hire data scientists amongst a group of candidates who have taken training courses hosted by the company. Many candidates sign up for training and the firm would like to know which trainees would be interested in working for the company and which are currently looking for new employment. Targeting these candidates reduces costs and time associated with hiring data scientists. In addition, a model of this kind would also help improve the quality of these trainings and the planning of courses that might further help categorize candidates in the future. The desired outcome of this model is to accurately predict the probability that a given trainee is on the job market and is actively looking for a new position. The ultimate action taken as a result of this ADS is determining which trainees should be invited to interview for data science positions at the company.

In the course of our analysis, we will examine these data, the cleaning process, and the system itself in order to evaluate whether the ADS discriminates on the basis of gender or socioeconomic status. If present, such discrimination could result in certain groups of employees receiving less rigorous training, fewer promotions and denied access to other benefits. Our aim is to design a “nutritional label” for the ADS that examines the bias in the data, the processing, and one machine learning model used in this decisionmaking process.

3 Input and Output

These data, made available by the company, are designed to understand the factors that lead a data scientist to search for a new job. Important information pertaining to demographics, education and prior experience is made available (on an anonymized basis) from registered and enrolled candidates. There is little information available about the geography of the candidate population in question or even the time during which these data were collected however we can discern from that they represent job candidates from 123 different cities of varying levels of development. The training dataset for the ADS comprises 19,158 candidates with the following 14 features describing each:

Feature	Description
enrollee_id	Unique ID for candidate
city	Code given for candidate's city
city_development_index	Development index of the city (scaled)
gender	Gender of candidate
relevant_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate total experience in years
company_size	Number of employees in current employer's company
company_type	Type of current employer
last_new_job	Difference in years between previous job and current job
training_hours	training hours completed
target	0 - Not looking for job change, 1 – Looking for a job change

The majority of these features are categorical(Nominal, Ordinal, Binary), some with high cardinality. The exceptions are "city_development_index" and "training hours" both of which are numeric. Other variables such as "experience" and "company_size" which could potentially be numeric, are presented as aggregated categorical variables such as '50-99', '<10', '10000+', '5000-9999', '1000-4999', '10-49', '100-500', and '500-999'.

Some features have large proportions of missing values, particularly "gender", "company_type" and 'company_size".

In terms of relative distributions, this dataset is largely imbalanced on our target variable with 75% of candidates not seeking a new job and only 25% actively on the job market. Discuss distributions.....

- Describe the data used by this ADS. How was this data collected or selected?
- For each input feature, describe its datatype, give information on missing values and on the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting or appropriate
- What is the output of the system (e.g. is it a class label, a score, a probability or some other type of output) and how do we interpret it?

The output of this ADS comes from a logistic regression model that was trained on a synthetically upsampled data set to correct for imbalance in the target variable. As a result, the output can be

Feature	Data Type	Distinct Values	% Null
enrollee_id	int64	19,158	0.0 %
city	object	123	0.0 %
city_development_index	float64	93	0.0 %
gender	object	4	23.5 %
relevant_experience	object	2	0.0 %
enrolled_university	object	4	2.0 %
education_level	object	6	2.4 %
major_discipline	object	7	14.7 %
experience	object	23	0.7 %
company_size	object	9	31.0 %
company_type	object	7	32.0 %
last_new_job	object	7	2.2 %
training_hours	int64	241	0.0 %
target	int64	2	0.0 %

interpreted as a numeric value between 0 and 1 that represents the probability of a candidate being on the job market (or of having a target value equal to 1).

4 Implementation and Validation

The ultimate model used for this ADS is a logistic regression trained on a dataset that was augmented using the Synthetic Minority Oversampling Technique (SMOTE). Prior to training the model however, the creator made several choices pertaining to the preparation and pre-processing of the data that had an ultimate impact on its final results.

1. data cleaning
2. SMOTE
3. models that were trained
4. regularization/hyper parameter tuning if any
5. how model was validated - against what dataset
6. measured performance on following metrics

Present your understanding of the code that implements this ADS. This code was implemented by others in this part of the assignment. Your goal here is to demonstrate that you understand the implementation at a high level.

- Describe data cleaning and any other pre-processing
- Give high-level information about the implementation of the system
- How was the ADS validated? How do we know that it meets its stated goal(s)?

5 Outcomes

- Analyze the effectiveness (accuracy) of the ADS by comparing it's performance across subpopulations
- Select one or more fairness or diversity measures, justify your choice of these measures for the ADS in question and quantify the fairness or diversity of this ADS.
- Develop additional methods for analyzing ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME) or any other property that you believe is important to check for this ADS

6 Conclusion & Summary

- Do you believe that the data was appropriate for this ADS?
- Do you believe the implementation is robust, accurate and fair? Discuss any choice of accuracy and fairness measures and explain which stakeholders may find these measures appropriate.
- Would you be comfortable deploying this ADS in the public sector, or in industry? Why so or why not?
- What improvements do you recommend to the data collection, processing or analysis methodology?