Oleksandr Romanko, Ph.D.
Senior Research Analyst, Risk Analytics, Watson Financial Services, IBM Canada
Adjunct Professor, University of Toronto

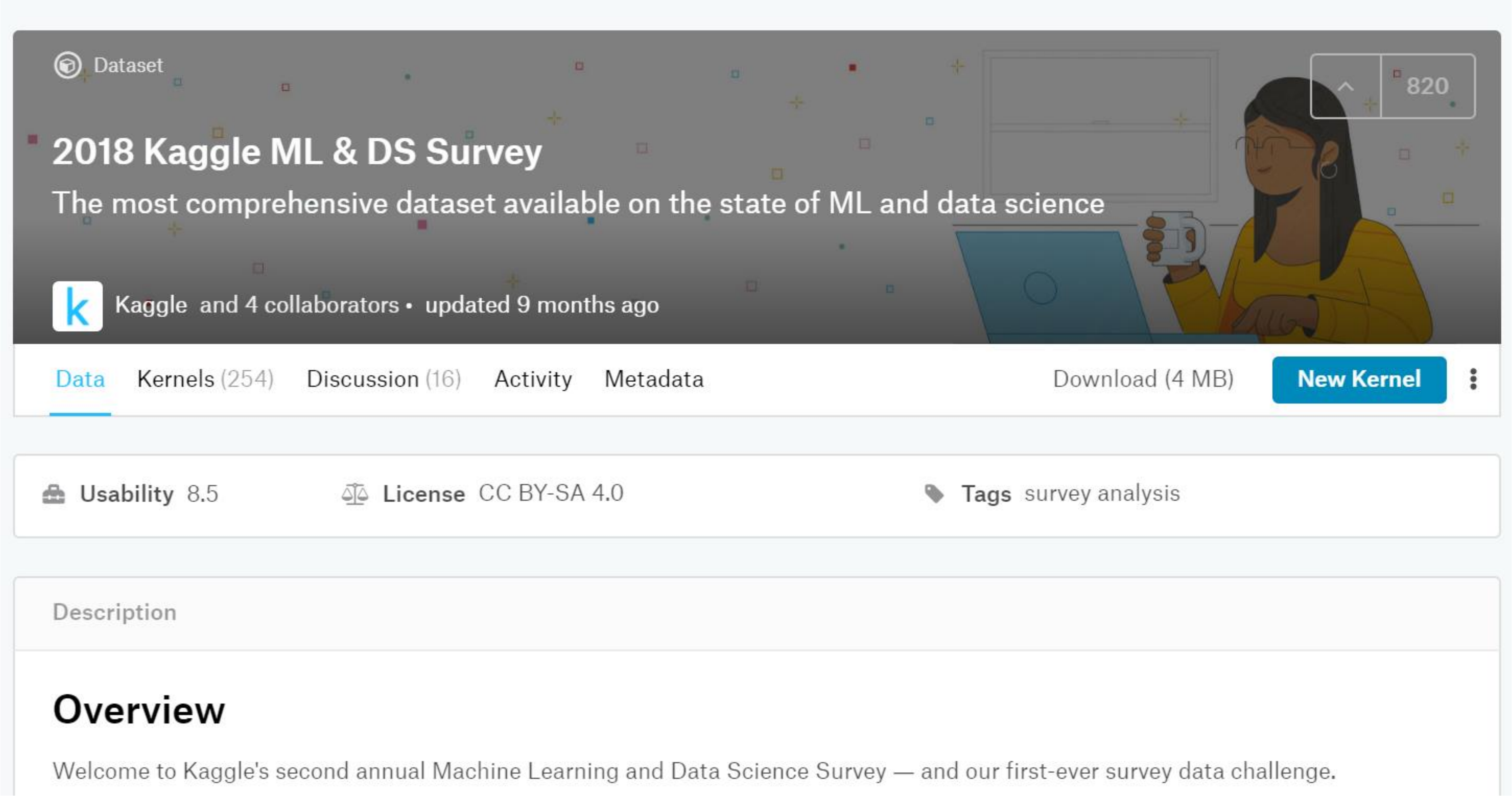
Xinrun Shan
Teaching Assistant, University of Toronto

MIE1624H – Introduction to Data Science and Analytics

Assignment 1 – Salary Classification Problem

University of Toronto
October 7, 2019

Kaggle ML & DS Survey 2018 – https://www.kaggle.com/kaggle/kaggle-survey-2018/



2018 Kaggle ML & DS Survey

The most comprehensive dataset available on the state of ML and data science

Kaggle and 4 collaborators • updated 9 months ago

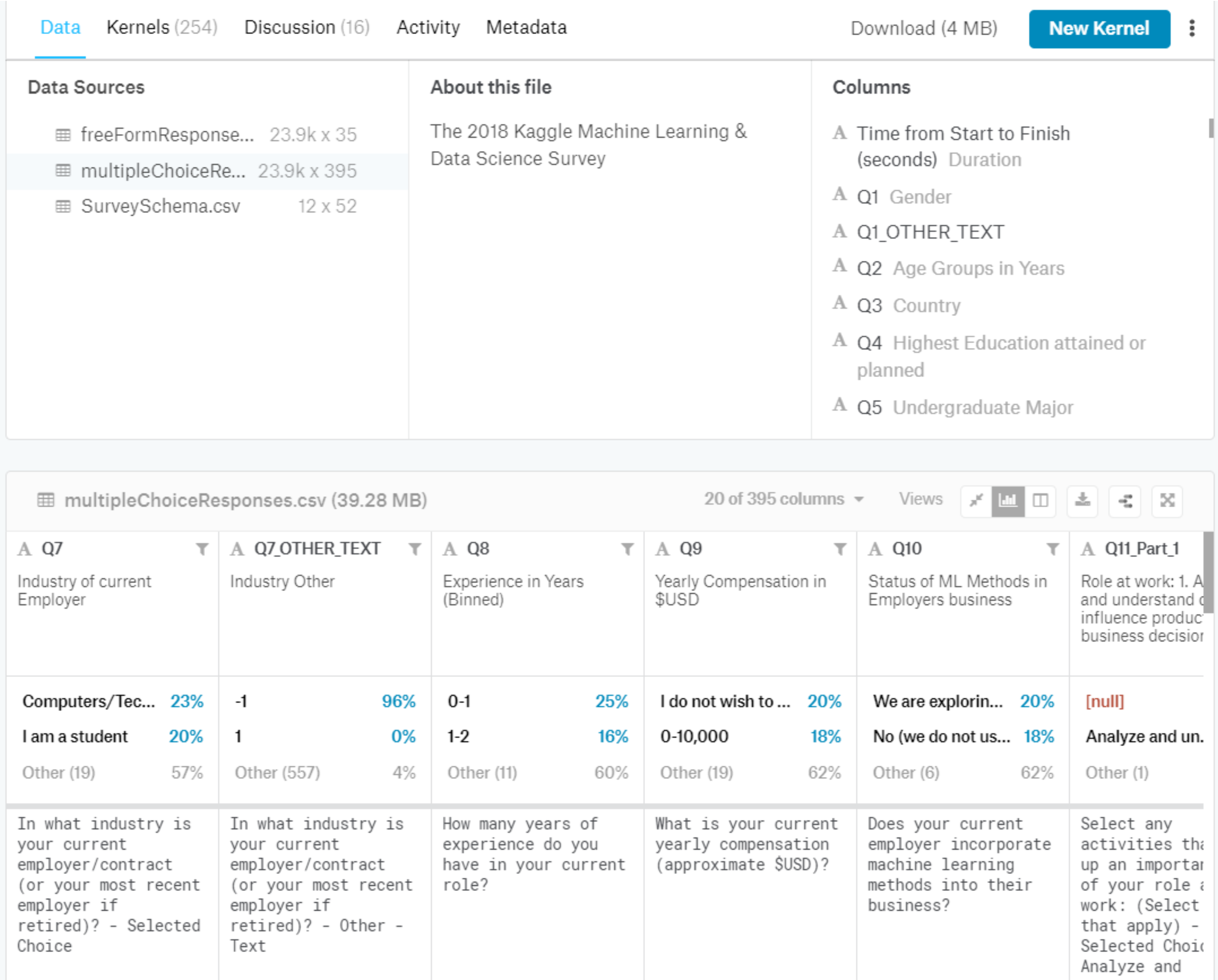
Usability 8.5 License CC BY-SA 4.0 Tags survey analysis

Description

Overview

Welcome to Kaggle's second annual Machine Learning and Data Science Survey — and our first-ever survey data challenge.

Kaggle ML & DS Survey 2018 – https://www.kaggle.com/kaggle/kaggle-survey-2018/



Data Sources

freeFormResponse...	23.9k x 35
multipleChoiceRe...	23.9k x 395
SurveySchema.csv	12 x 52

About this file

The 2018 Kaggle Machine Learning & Data Science Survey

Columns

- Time from Start to Finish (seconds) Duration
- Q1 Gender
- Q1_OTHER_TEXT
- Q2 Age Groups in Years
- Q3 Country
- Q4 Highest Education attained or planned
- Q5 Undergraduate Major

multipleChoiceResponses.csv (39.28 MB)

Q7		Q7_OTHER_TEXT		Q8		Q9		Q10		Q11_Part_1
Industry of current Employer		Industry Other		Experience in Years (Binned)		Yearly Compensation in $USD		Status of ML Methods in Employers business		Role at work: 1. A and understand c influence produc business decisior
Computers/Tec...	23%	-1	96%	0-1	25%	I do not wish to ...	20%	We are explorin...	20%	[null]
I am a student	20%	1	0%	1-2	16%	0-10,000	18%	No (we do not us...	18%	Analyze and un.
Other (19)	57%	Other (557)	4%	Other (11)	60%	Other (19)	62%	Other (6)	62%	Other (1)
In what industry is your current employer/contract (or your most recent employer if retired)? - Selected Choice		In what industry is your current employer/contract (or your most recent employer if retired)? - Other - Text		How many years of experience do you have in your current role?		What is your current yearly compensation (approximate $USD)?		Does your current employer incorporate machine learning methods into their business?		Select any activities tha up an importar of your role a work: (Select that apply) - Selected Choic Analyze and

Assignment goal

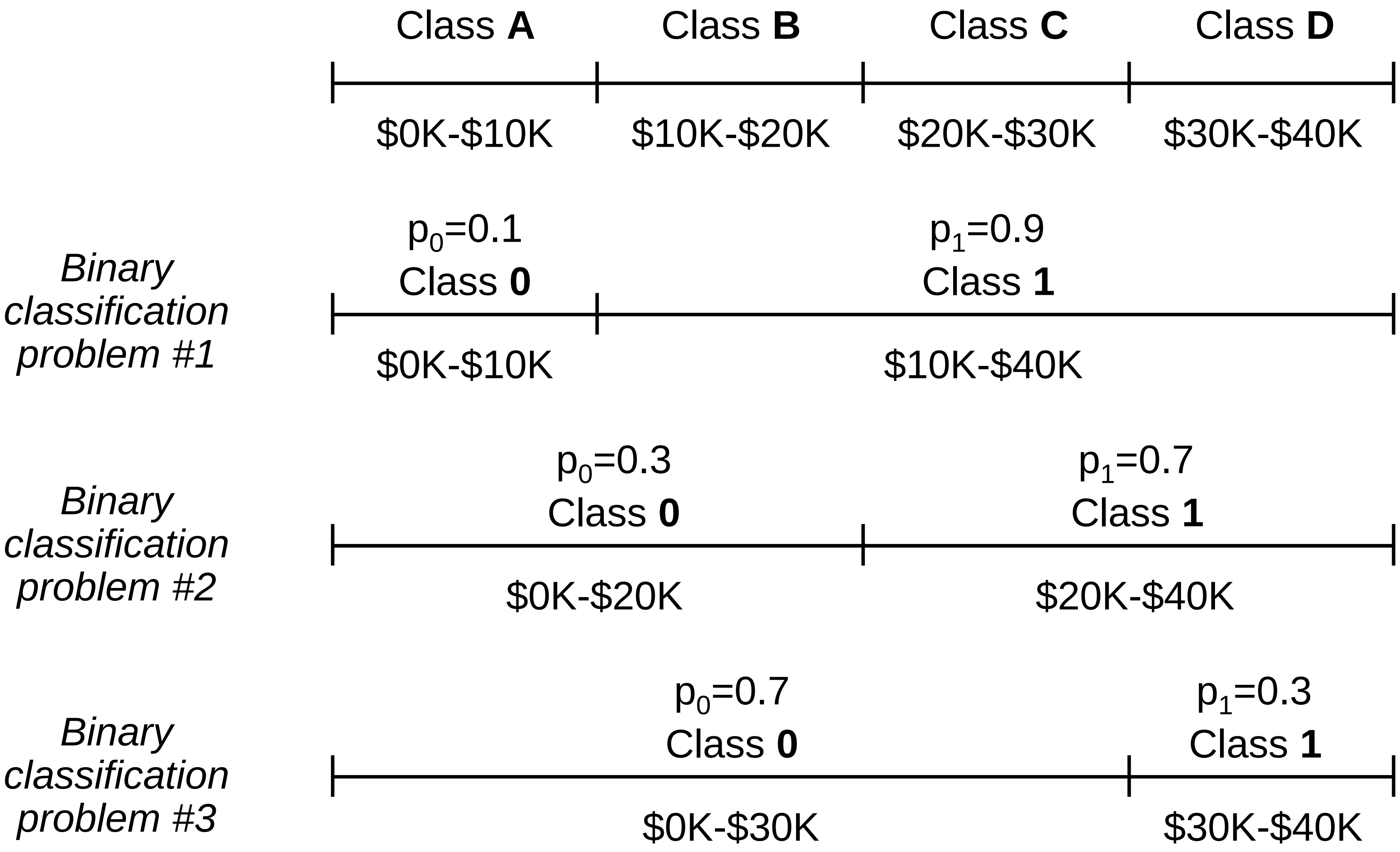
Based on "2018 Kaggle ML & DS Survey" **identify skills** that are **important for data scientists** to posses based on salary level as a target variable. The goal is to **identify skills and qualifications from the survey that are important to learn in data science field in order to have a high salary**. Supervised machine learning algorithm (logistic regression) would be used for the assignment. In your free time, you can also experiment with other supervised learning algorithms, e.g., decision trees, and unsupervised machine learning algorithms, e.g., clustering of skills.

Possible decisions to make based on modeling results (not part of the assignment):

- Design a study program at your university, e.g., Master of DS & AI
- Build a recommender system for taking online courses
- Design personalized study program that maximizes your salary
- Any other creative use of insights from the survey

If you are interested to continue with this type of project as your research project or MEng project, please e-mail the course instructor.

Ordinary multi-class classification algorithm



Class **A** Class **B** Class **C** Class **D**

\$0K-\$10K \$10K-\$20K \$20K-\$30K \$30K-\$40K

Binary classification problem #1

p_0=0.1
Class **0**
\$0K-\$10K

p_1=0.9
Class **1**
\$10K-\$40K

Binary classification problem #2

p_0=0.3
Class **0**
\$0K-\$20K

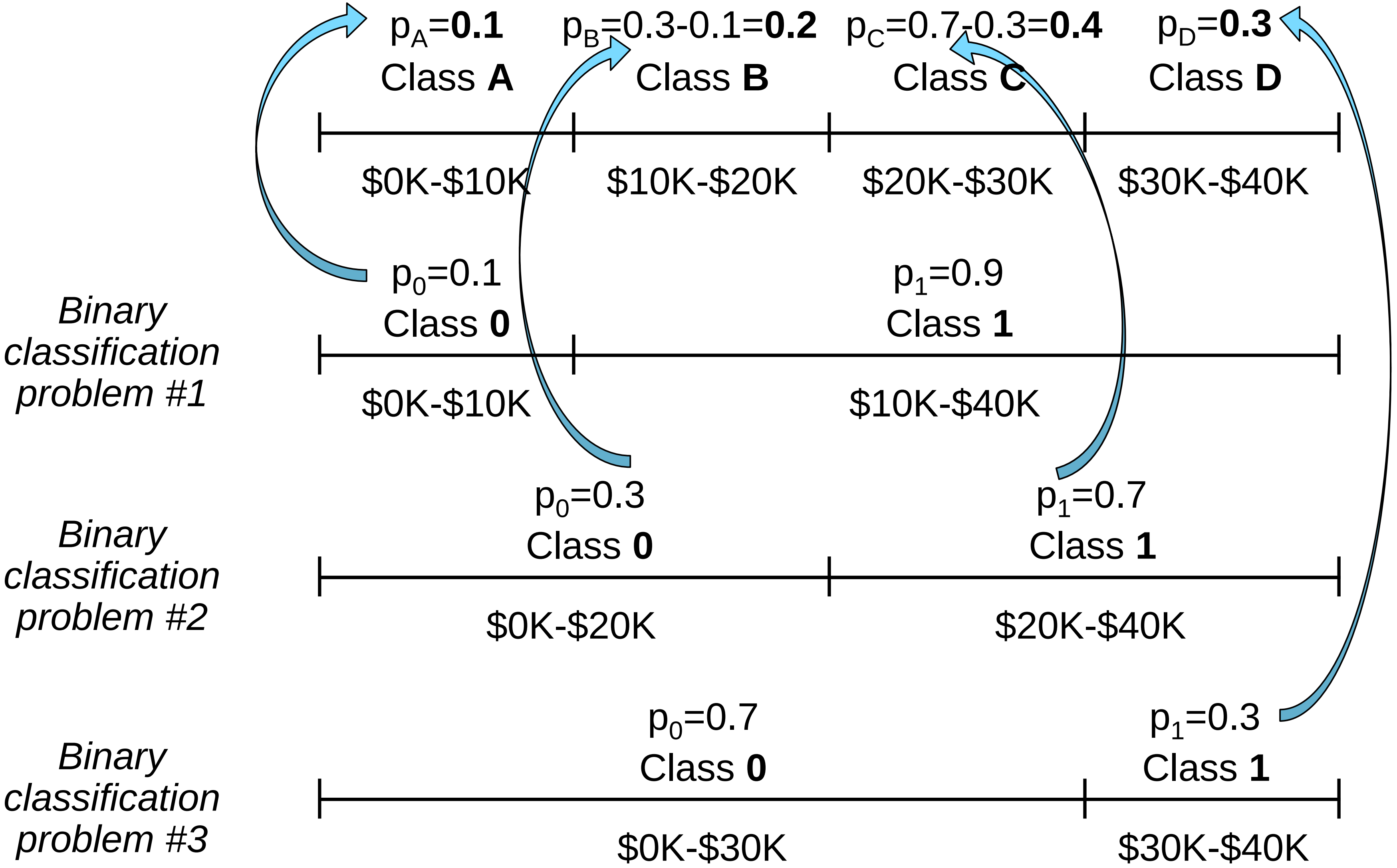
p_1=0.7
Class **1**
\$20K-\$40K

Binary classification problem #3

p_0=0.7
Class **0**
\$0K-\$30K

p_1=0.3
Class **1**
\$30K-\$40K

Ordinary multi-class classification algorithm



p_A=**0.1** Class **A** $0K-$10K

p_B=0.3-0.1=**0.2** Class **B** $10K-$20K

p_C=0.7-0.3=**0.4** Class **C** $20K-$30K

p_D=**0.3** Class **D** $30K-$40K

Binary classification problem #1

p_0=0.1 Class **0** $0K-$10K

p_1=0.9 Class **1** $10K-$40K

Binary classification problem #2

p_0=0.3 Class **0** $0K-$20K

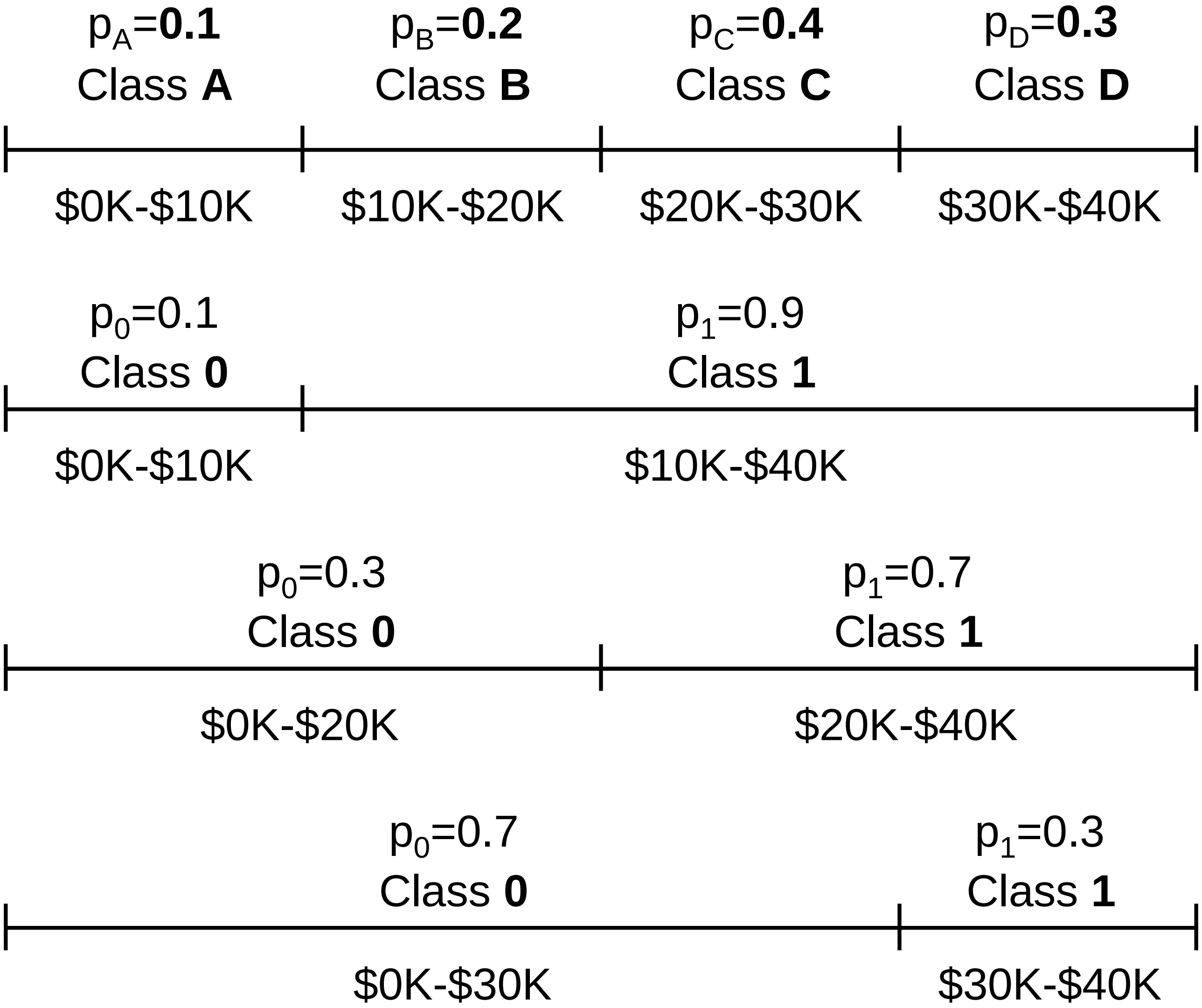
p_1=0.7 Class **1** $20K-$40K

Binary classification problem #3

p_0=0.7 Class **0** $0K-$30K

p_1=0.3 Class **1** $30K-$40K

Ordinary multi-class classification algorithm



p_A=**0.1** Class **A** p_B=**0.2** Class **B** p_C=**0.4** Class **C** p_D=**0.3** Class **D**

$0K-$10K $10K-$20K $20K-$30K $30K-$40K

Binary classification problem #1

p_0=0.1 Class **0** p_1=0.9 Class **1**

$0K-$10K $10K-$40K

Binary classification problem #2

p_0=0.3 Class **0** p_1=0.7 Class **1**

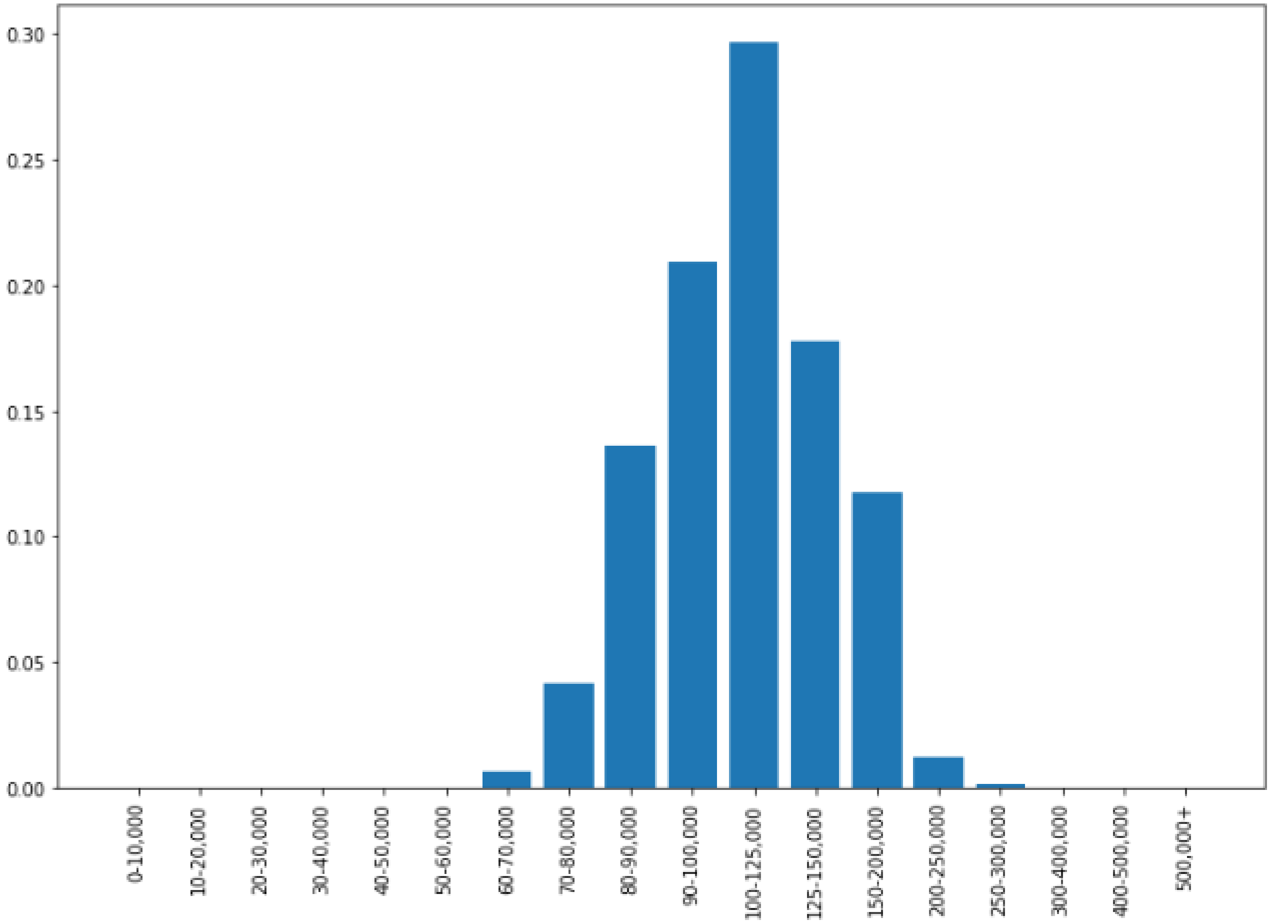
$0K-$20K $20K-$40K

Binary classification problem #3

p_0=0.7 Class **0** p_1=0.3 Class **1**

$0K-$30K $30K-$40K

Algorithm accuracy



$100-$125K 30% accuracy

$90-$150K 70% accuracy

$80-$200K 94% accuracy