JOB APPLICATION MARKET

GW Data Science

TEAM GLOBAL:

Sandhya Karki
Sajan Kumar Kar
Neelima Puthanveetil
Sai Pavan Kumar

GOALS

01	Understanding Employability Factors:	To gain a comprehensive understanding of the various factors that influence employability in the tech industry
02	Guidance for Job Seekers	Offer actionable advice to job seekers in the tech industry, helping them understand what factors are most likely to enhance their employability and career development.
03	Empower Companies and Recruiters:	Equip companies and recruiters with data-driven insights to develop more effective and tailored recruitment strategies.
04	Mental Health and Career Outcomes	Explore the potential impact of mental health on career progression and salary, particularly in the context of the tech industry.
05	Predicting employbility	Utilize data-driven models to accurately forecast job applicant success in the tech industry based on key employability factors.

DATASET OVERVIEW

Why did we choose this dataset?

Relevance

Real-World Application:

Provides real-world data from over 70,000 directly aligning with our research domain.

Diverse Metrics:

Includes essential metrics like employment status, education level, and more, offering critical insights for our study.

Data Quality

Well-Structured Format:

The data is well-organized, facilitating easy access and interpretation.

Easy Cleaning and Pre-processing:

Demonstrates high data cleanliness and standardization, simplifying the pre-processing steps.

DATASET OVERVIEW

About the Dataset

73,000+
records

15 columns in total

7
integer-type columns

index, employment, yearscode, yearscodepro and previoussalary, computerskills, employed

8
char-type columns

age, accessibulity, edLevel, gender, mentalHealth, mainBranch, country, haveWorkedWith,

FEATURES OF THE DATASET

Index

the count/unique identifier for each record (1, 2, 3 ...) (integer)

Edlevel

education level of the applicant (Undergraduate, Master, PhD...) *(categorical)*

Age

age of the applicant, >35 years old or <35 years old *(categorical)*

Gender

gender of the applicant, (Man, Woman, or NonBinary) *(categorical)*

PreviousSalary

the applicant's previous job salary *(float)*

ComputerSkills

number of computer skills known by the applicant *(integer)*

Employed

target variable, whether the applicant has been hired *(categorical)*

Some of the type of columns and their description (page 1)

FEATURES OF THE DATASET

HaveWorkedWith

the technical skills of the applicant (C++, Python, Git....)(string/ categorical)

MainBranch

whether the applicant is a professional developer *(categorical)*

YearsCode

how long the applicant has been coding *(integer)*

Accessibility

if the applicant has any accessibility needs. (Yes or No) (categorical)

Employment

if the applicant is currently employed (0 meaning No and 1 meaning Yes) (integer)

Country

the country of the applicant. (US, Canada etc) (categorical)

YearsCodePro

how long the applicant has been coding in a professional context, *(integer)*

Mental Health

if the applicant have any mental health issues. denoted by "Yes" or "No" (integer)

Some of the type of columns and their description (page 2)

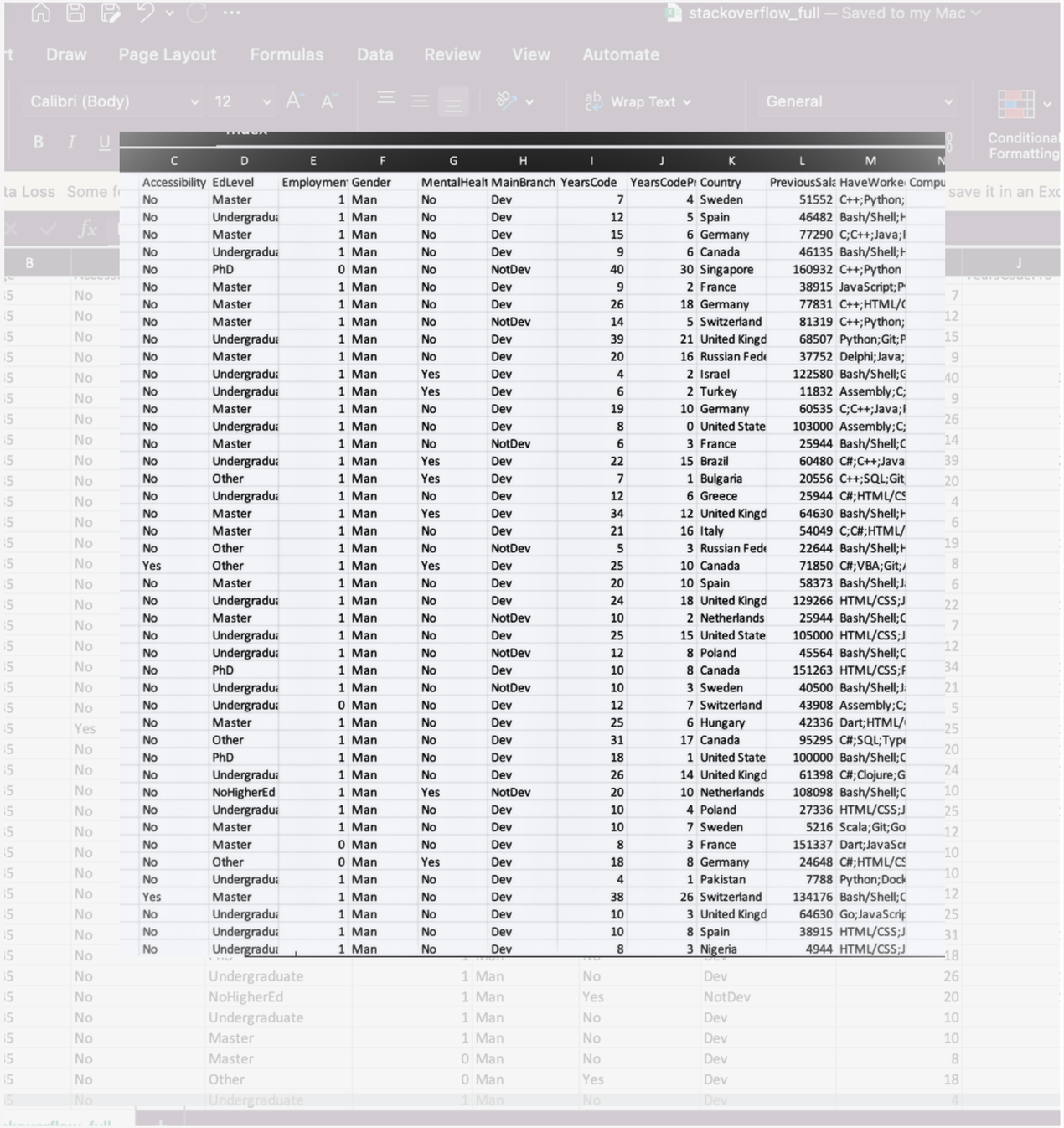
DATA CLEANING

Missing and Duplicate Check

Numeric Feature Processing

Character Feature Processing

Outlier Detection and Removal



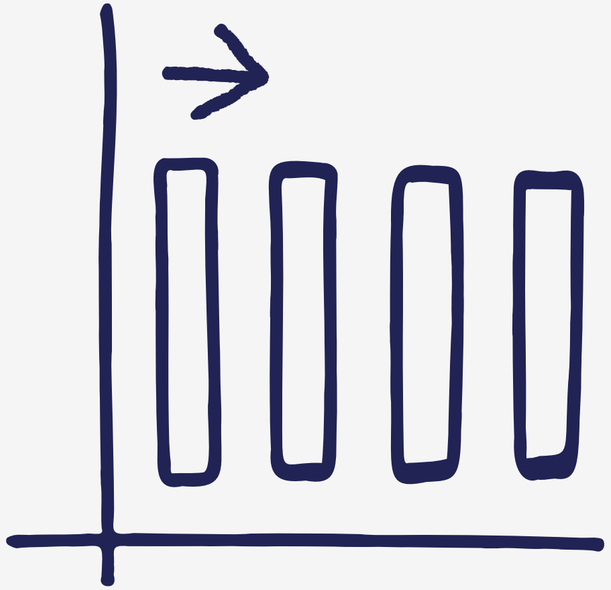
Accessibility	EdLevel	Employmen	Gender	MentalHealt	MainBranch	YearsCode	YearsCodePr	Country	PreviousSala	HaveWorke	Compu
No	Master	1	Man	No	Dev	7	4	Sweden	51552	C++;Python;	
No	Undergradua	1	Man	No	Dev	12	5	Spain	46482	Bash/Shell;H	
No	Master	1	Man	No	Dev	15	6	Germany	77290	C;C++;Java;I	
No	Undergradua	1	Man	No	Dev	9	6	Canada	46135	Bash/Shell;H	
No	PhD	0	Man	No	NotDev	40	30	Singapore	160932	C++;Python	
No	Master	1	Man	No	Dev	9	2	France	38915	JavaScript;P	
No	Master	1	Man	No	Dev	26	18	Germany	77831	C++;HTML/(	
No	Master	1	Man	No	NotDev	14	5	Switzerland	81319	C++;Python;	
No	Undergradua	1	Man	No	Dev	39	21	United Kingd	68507	Python;Git;P	
No	Master	1	Man	No	Dev	20	16	Russian Fede	37752	Delphi;Java;	
No	Undergradua	1	Man	Yes	Dev	4	2	Israel	122580	Bash/Shell;C	
No	Undergradua	1	Man	Yes	Dev	6	2	Turkey	11832	Assembly;C;	
No	Master	1	Man	No	Dev	19	10	Germany	60535	C;C++;Java;I	
No	Undergradua	1	Man	No	Dev	8	0	United State	103000	Assembly;C;	
No	Master	1	Man	No	NotDev	6	3	France	25944	Bash/Shell;C	
No	Undergradua	1	Man	Yes	Dev	22	15	Brazil	60480	C#;C++;Java	
No	Other	1	Man	Yes	Dev	7	1	Bulgaria	20556	C++;SQL;Git	
No	Undergradua	1	Man	No	Dev	12	6	Greece	25944	C#;HTML/CS	
No	Master	1	Man	Yes	Dev	34	12	United Kingd	64630	Bash/Shell;H	
No	Master	1	Man	No	Dev	21	16	Italy	54049	C;C#;HTML/	
No	Other	1	Man	No	NotDev	5	3	Russian Fede	22644	Bash/Shell;H	
Yes	Other	1	Man	Yes	Dev	25	10	Canada	71850	C#;VBA;Git;/	
No	Master	1	Man	No	Dev	20	10	Spain	58373	Bash/Shell;J	
No	Undergradua	1	Man	No	Dev	24	18	United Kingd	129266	HTML/CSS;J	
No	Master	1	Man	No	NotDev	10	2	Netherlands	25944	Bash/Shell;C	
No	Undergradua	1	Man	No	Dev	25	15	United State	105000	HTML/CSS;J	
No	Undergradua	1	Man	No	NotDev	12	8	Poland	45564	Bash/Shell;C	
No	PhD	1	Man	No	Dev	10	8	Canada	151263	HTML/CSS;F	
No	Undergradua	1	Man	No	NotDev	10	3	Sweden	40500	Bash/Shell;J	
No	Undergradua	0	Man	No	Dev	12	7	Switzerland	43908	Assembly;C;	
No	Master	1	Man	No	Dev	25	6	Hungary	42336	Dart;HTML/	
No	Other	1	Man	No	Dev	31	17	Canada	95295	C#;SQL;Type	
No	PhD	1	Man	No	Dev	18	1	United State	100000	Bash/Shell;C	
No	Undergradua	1	Man	No	Dev	26	14	United Kingd	61398	C#;Clojure;G	
No	NoHigherEd	1	Man	Yes	NotDev	20	10	Netherlands	108098	Bash/Shell;C	
No	Undergradua	1	Man	No	Dev	10	4	Poland	27336	HTML/CSS;J	
No	Master	1	Man	No	Dev	10	7	Sweden	5216	Scala;Git;Go	
No	Master	0	Man	No	Dev	8	3	France	151337	Dart;JavaScr	
No	Other	0	Man	Yes	Dev	18	8	Germany	24648	C#;HTML/CS	
No	Undergradua	1	Man	No	Dev	4	1	Pakistan	7788	Python;Dock	
Yes	Master	1	Man	No	Dev	38	26	Switzerland	134176	Bash/Shell;C	
No	Undergradua	1	Man	No	Dev	10	3	United Kingd	64630	Go;JavaScrip	
No	Undergradua	1	Man	No	Dev	10	8	Spain	38915	HTML/CSS;J	
No	Undergradua	1	Man	No	Dev	8	3	Nigeria	4944	HTML/CSS;J	

ALL MISSING RECORDS	**0**
DUPLICATE RECORDS	**0**

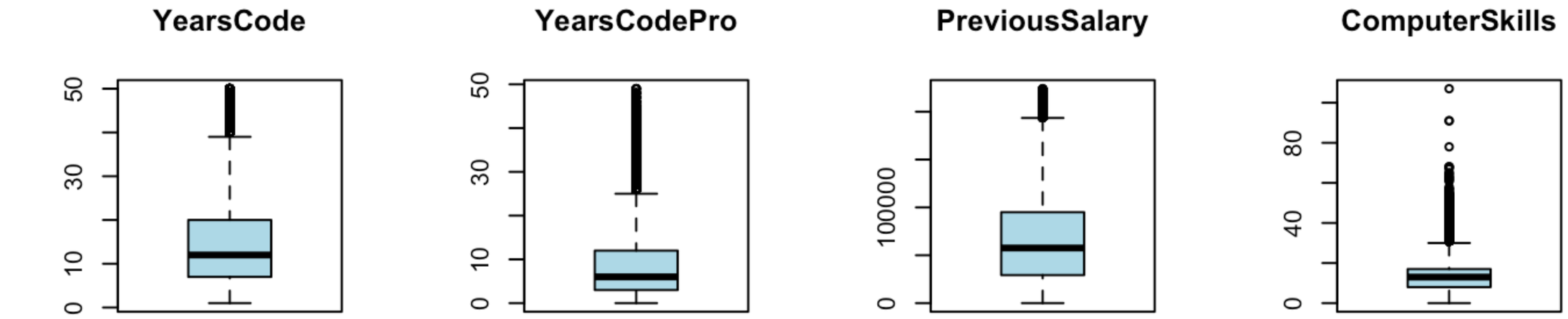
Description: df [6 × 7]

	X <int>	Employment <int>	YearsCode <int>	Employed <int>	PreviousSalary <dbl>	ComputerSkills <int>
1	0	1	7	0	51552	4
2	1	1	12	1	46482	12
3	2	1	15	0	77290	7
4	3	1	9	0	46135	13
5	4	0	40	0	160932	2
6	5	1	9	0	38915	5

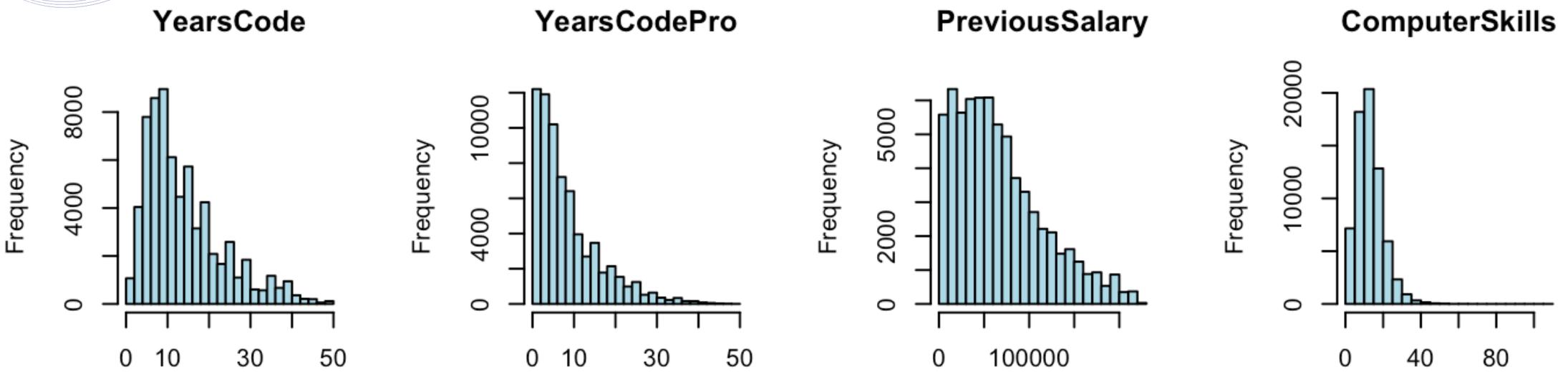
- ***The provided R code transforms the "Employment" and "Employed" columns in the dataframe***
- ***"df_clean." It converts numeric values to descriptive categories, enhancing interpretability for analysis.***
- ***The resulting factors are "currently_employed" or "not_currently_employed" for "Employment" and "hired" or "not_hired" for "Employed."***



if there is data where their professional coding experience is more than their coding experience, we have to remove it because that's faulty data.


YearsCode
YearsCodePro
PreviousSalary
ComputerSkills
0
10
30
50
100000
40
80

OUTLIERS


YearsCode
YearsCodePro
PreviousSalary
ComputerSkills
Frequency
0
4000
8000
10000
2000
5000
20000
10
30
50
100000
40
80

E D A

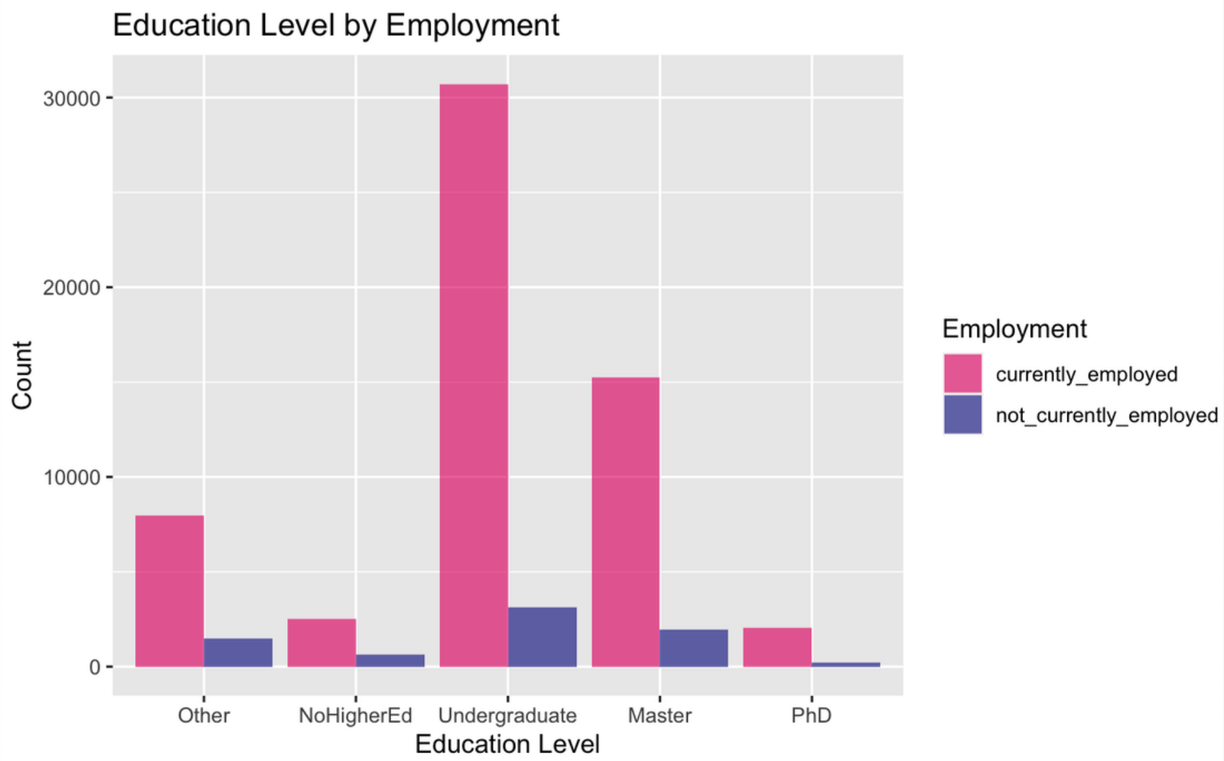
Exploratory data analysis

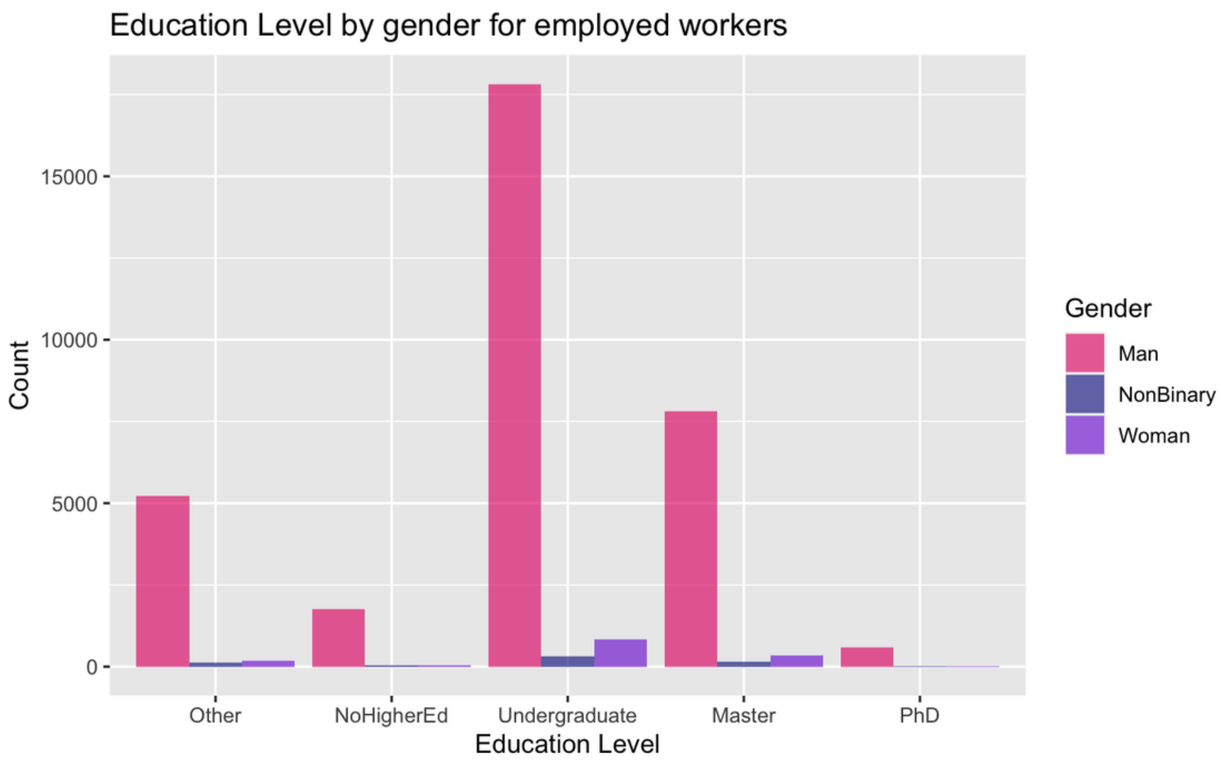
+

HYPOTHESIS TESTING

SMART Q1:

Is there a significant difference between employed males and non-males in their respective education levels?


Education Level by Employment
30000
20000
10000
0
Count
Other
NoHigherEd
Undergraduate
Master
PhD
Education Level
Employment
currently_employed
not_currently_employed


Education Level by gender for employed workers
15000
10000
5000
0
Count
Other
NoHigherEd
Undergraduate
Master
PhD
Education Level
Gender
Man
NonBinary
Woman

SMART Q1: HYPOTHESIS TESTING

TWO PORPORTIONS Z TEST

H0: THERE IS **NO DIFFERENCE** IN THE NUMBER OF **UNDERGRADUATE** MALES WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

HA: THERE **IS A SIGNIFICANT DIFFERENCE** IN THE NUMBER OF **UNDERGRADUATE** MALES WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

.....

90.72% OF MALES WITH **BACHELOR'S** DEGREES WERE EMPLOYED

91.39% OF NON-MALES WITH **BACHELOR'S** DEGREES WERE EMPLOYED

P-VALUE: .2951

H0: THERE IS **NO DIFFERENCE** IN THE NUMBER OF MALES WITH A **MASTER'S DEGREE** WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

HA: THERE **IS A SIGNIFICANT DIFFERENCE** IN THE NUMBER OF MALES WITH A **MASTER'S DEGREE** WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

.....

88.45% OF MALES WITH **MASTER'S** DEGREES WERE EMPLOYED

44.46% OF NON-MALES WITH **MASTERS'S** DEGREES WERE EMPLOYED

P-VALUE: < 2.2X10^-16

H0: THERE IS **NO DIFFERENCE** IN THE NUMBER OF MALES WITH A **PHD** WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

HA: THERE **IS A SIGNIFICANT DIFFERENCE** IN THE NUMBER OF MALES WITH A **PHD** WHO ARE EMPLOYED WHEN COMPARED TO NONMALES

.....

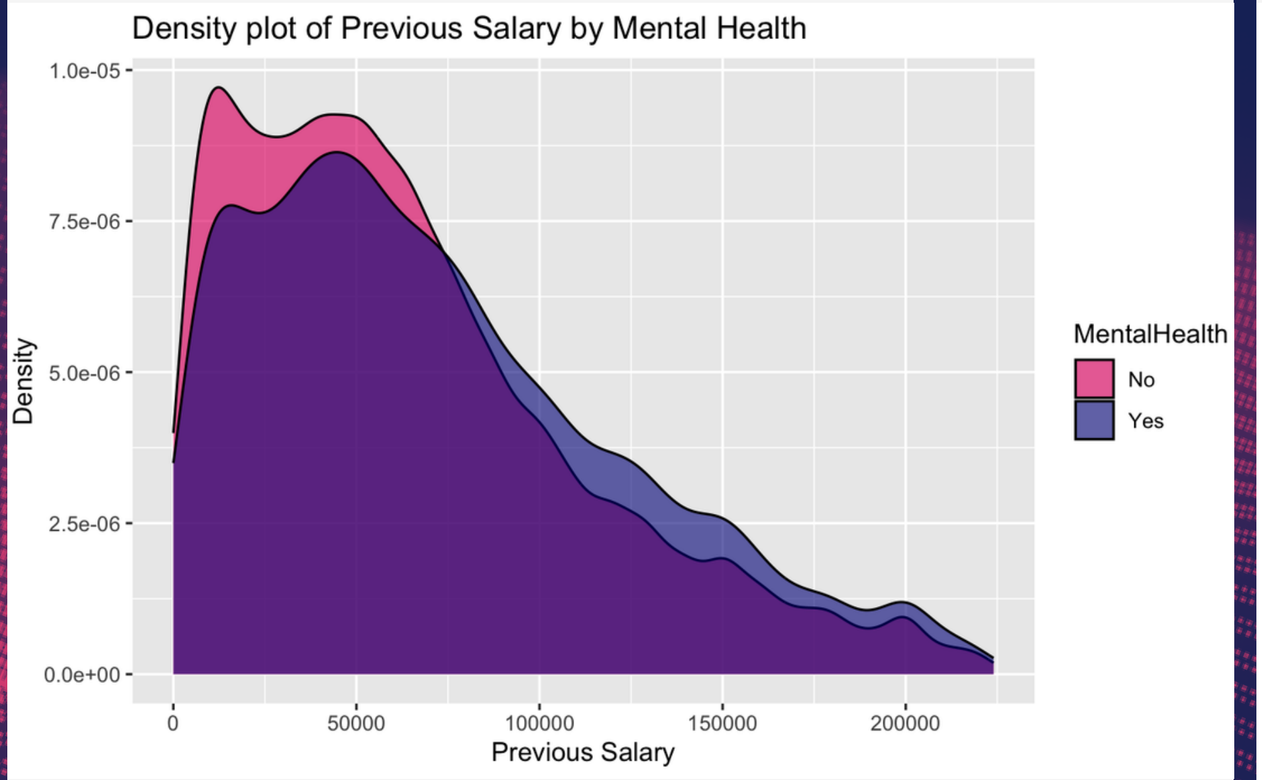
90.64% OF MALES WITH **PHD** DEGREES WERE EMPLOYED

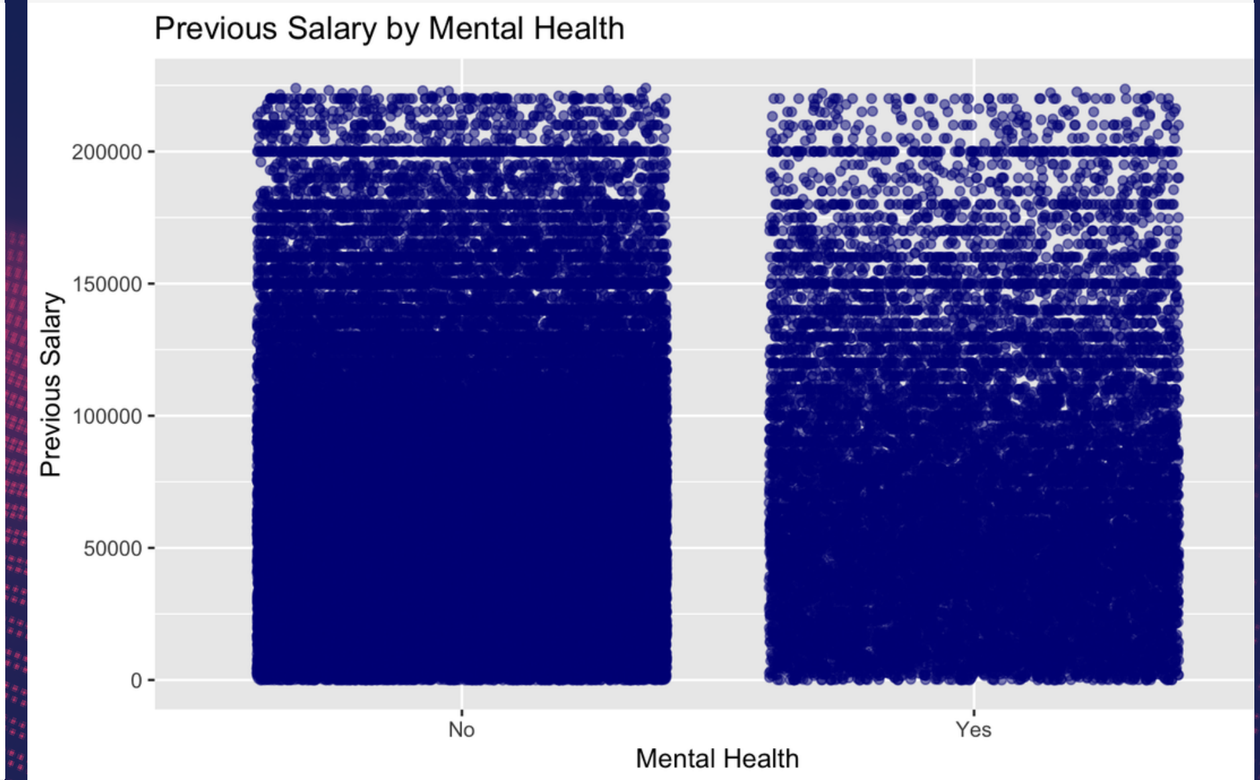
89.50% OF NON-MALES WITH **PHD** DEGREES WERE EMPLOYED

P-VALUE: .7119

SMART Q2:

Is an individual's mental health associated with their previous salary?


Density plot of Previous Salary by Mental Health
1.0e-05
7.5e-06
5.0e-06
2.5e-06
0.0e+00
Density
0
50000
100000
150000
200000
Previous Salary
MentalHealth
No
Yes


Previous Salary by Mental Health
200000
150000
100000
50000
0
Previous Salary
No
Yes
Mental Health

SMART Q2: HYPOTHESIS TESTING

TWO SAMPLE T TEST FOR MEANS

H0: THERE IS **NO DIFFERENCE** IN MEAN SALARY BETWEEN INDIVIDUALS WHO HAVE AND DO NOT HAVE MENTAL HEALTH ISSUES

HA: THERE IS **A SIGNIFICANT DIFFERENCE** IN MEAN SALARY BETWEEN INDIVIDUALS WHO HAVE AND DO NOT HAVE MENTAL HEALTH ISSUES

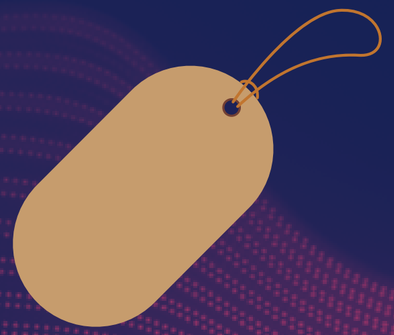
THOSE WHO RESPONDED **YES** TO HAVING MENTAL HEALTH ISSUES HAD A MEAN SALARY OF **$72,709**

THOSE WHO RESPONDED **NO** TO HAVING MENTAL HEALTH ISSUES HAD A MEAN SALARY OF **$64,652**

P-VALUE: < 2.2X10^-16

MODEL BUILDING

DATA PREP: ENCODING



- MAIN BRANCH
- MENTAL HEALTH
- ACCESSIBILITY
- EMPLOYMENT
- EDLEVEL
- AGE

- COUNTRY
- HAVEWORKEDWITH

- GENDER

[YearsCode, YearsCodePro, PreviosSalary, ComputerSkills]
The numerical features were standardized. No class imbalance in target variable.

MODELS

LOGISTIC REGRESSION

3-NEAREST NEIGHBOURS

SUPPORT VECTOR MACHINE

DECISION TREE CLASSIFIER

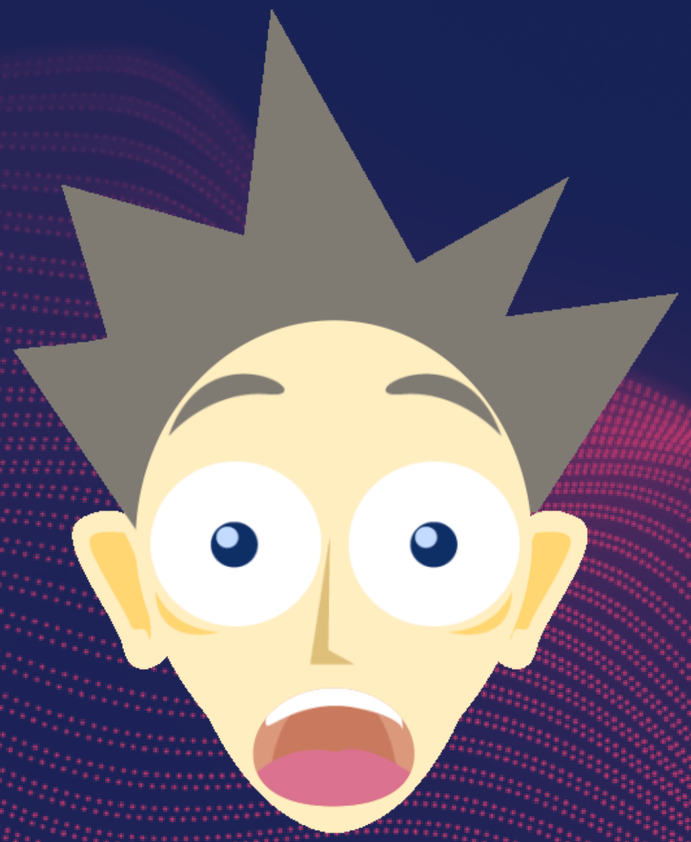
AUC-ROC

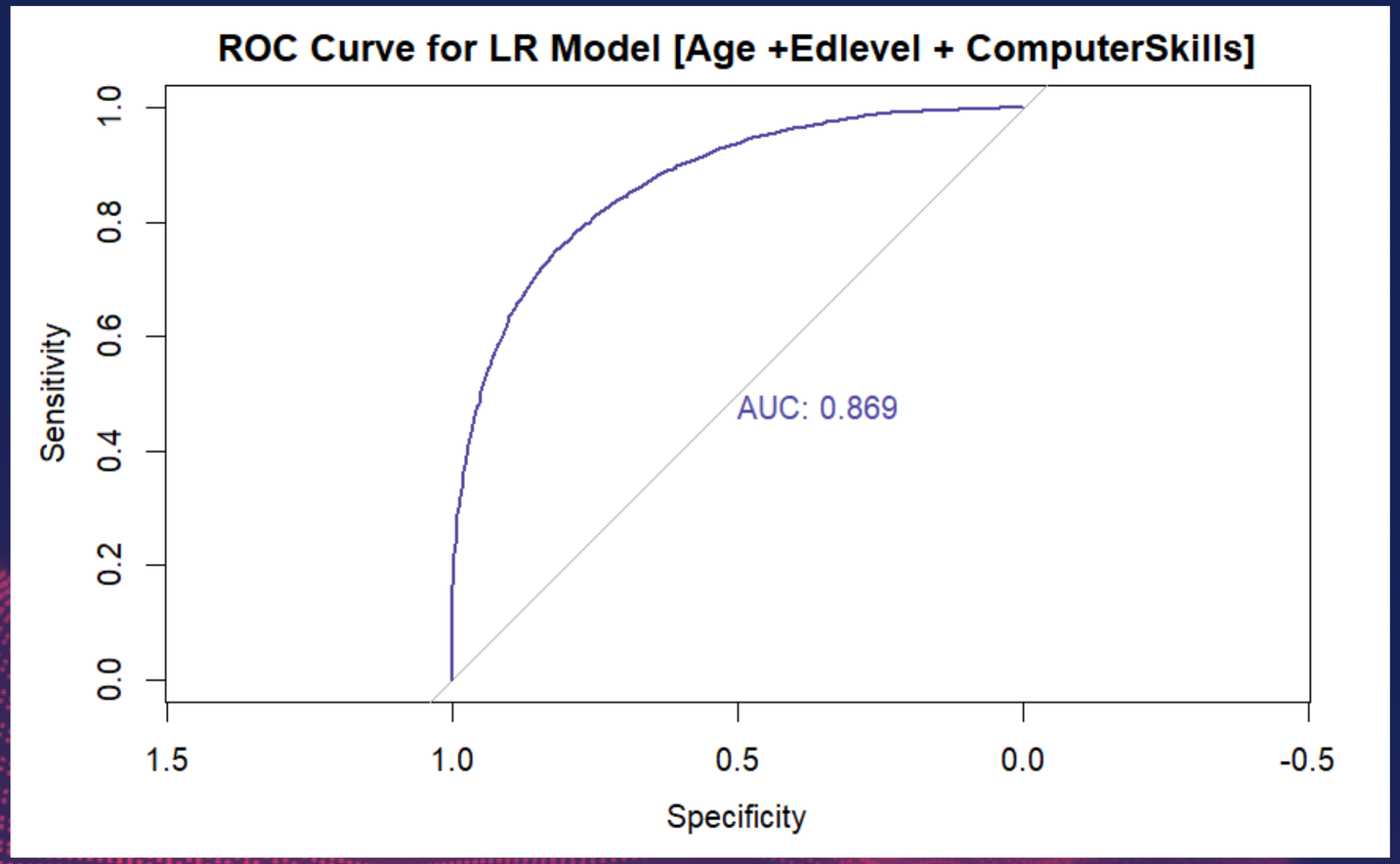
LR:	87.2%
3-NN:	49.6%
SVM:	78.4%
DT:	78.3%

Oddly, Decision Tree determined based on just `ComputerSkills`

SMART Q3:

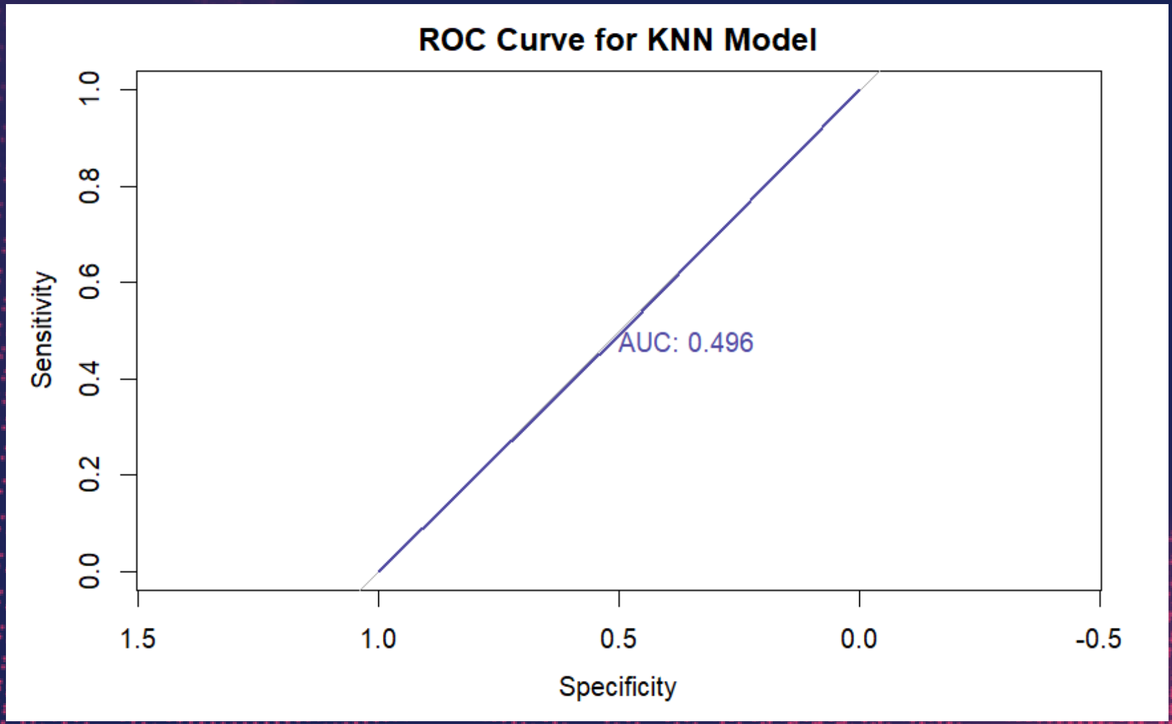
Can we predict with at least 60% accuracy whether an individual will be employed or not, based on just their Age, Education Level, and, Number of skills?

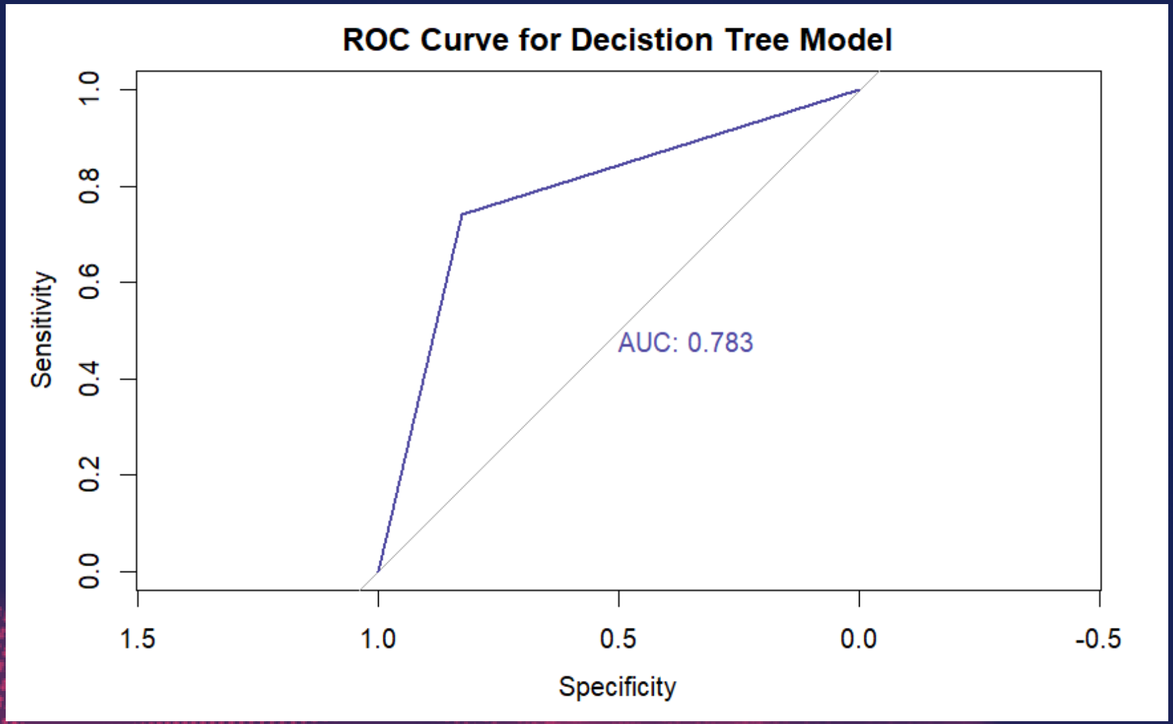



ROC Curve for LR Model [Age +Edlevel + ComputerSkills]
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.869
1.5
1.0
0.5
0.0
-0.5
Specificity

SMART Q4:

Is there a notable difference between distance-based models and tree-based models?


ROC Curve for KNN Model
Sensitivity
Specificity
AUC: 0.496
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5


ROC Curve for Decistion Tree Model
Sensitivity
Specificity
AUC: 0.783
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5

CONCLUSIONS

Come to us!! We will tell you if you'll get a job or not?

Consultation Fees: $5.99 😉

THANK YOU!

GW Data Science