MSc in Data Analytics

Advanced Data Mining

Foundation-Level Questions

F1: Appropriately encode the data. Remove correlated variables (if appropriate).

F2: Appropriately handle missing values: impute if you can, otherwise remove.

F3: Process and convert variables as necessary

F4: Prepare samples that are appropriate for building and evaluating machine learning models.

Basic Questions:

B1: Compute basic descriptive statistics of 4 independent variables

B2: Make 2 visualisations of the dependent variable against at least 2 other variables – comment on the findings

B3: Build a simple model that provides a reasonable benchmark of problem complexity – discuss your rationale.

B4: Discuss the class (im)balance in the dataset

Intermediate Questions

I1: Compare the performance of a Random Forest to the benchmark (naïve) model.

I2: Compare the performance of Naïve Bayes to the benchmark (naïve) model.

I3: Compare the performance of kNN (you must use both numeric and categorical variables, if you have them) to a benchmark model.

I4: Demonstrate overfitting on a model of your choice. (do not do this if you do I5)

I5: Build a model that demonstrates some aspect of the bias variance trade off – explain your answer (do not do this if you do I4)

I6: You suddenly recover the lost columns, redo Intermediate 1, 2, or 3 and discuss any observed differences

Advanced Questions

A1: Build an ensemble with 3 base models in it, and evaluate its performance against any other model(s) built

A2: Compare an agglomerative clustering algorithm with the class distribution, would you agree with the cluster assignment?

A3: Comprehensively demonstrate which is better: a C5.0, a Random Forest, or always picking the dominant class? **Don't use only the default settings!**

A4: Will a Support Vector Machine with a linear or polynomial kernel do better on this dataset?

A5: Does over or under sampling work better handling class imbalance? Use at 2 models to answer this question.

Show Off Questions

S1: Do some form of dimensionality reduction and demonstrate an improvement with/without it on Intermediate 1, 2, or 3. Note that PCA is not an option here.

S2: Using the full dataset, perform feature-engineering and/or feature selection. Demonstrate an improvement on any other model you have built, including those in S3 if you choose to do that question. Do not use Random Forests to perform feature selection.

S3: Using the full dataset, automate the hyperparameter optimisation of a Support Vector Machine or Neural Network. Demonstrate that you have not overfitted, but that it outperforms at least 2 other models you have built. **(Be careful to not commit too much time to this question!!)** Some credit will still be awarded if you don't beat 2 models.

Problem Overview

PASSNYC is a not-for-profit organization that facilitates a collective impact that is dedicated to broadening educational opportunities for New York City's talented and underserved students.

New York City is home to some of the most impressive educational institutions in the world, yet in recent years, the City's specialized high schools - institutions with historically transformative impact on student outcomes - have seen a shift toward more homogeneous student body demographics.

PASSNYC uses public data to identify students within New York City's under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of

students taking the Specialized High School Admissions Test (SHSAT). By focusing efforts in under-performing areas that are historically underrepresented in SHSAT registration, we will help pave the path to specialized high schools for a more diverse group of students.

Data Description

Column	Name	Description
1	Target	Level of economic need based on number of students in temp housing, HRA eligibility, free lunch eligibility. 3 levels – Low, Medium & High
2	Adjusted Grade	
3	New?	Is this record new
4	Other Location Code in LCGMS	
5	School Name	
6	SED Code	State Education Department
7	Location Code	
8	District	School District
9	Latitude	Geolocation
10	Longitude	Geolocation
11	Address (Full)	
12	City	
13	Zip	
14	Grades	The range of grade levels in this school
15	Grade Low	Lowest grade level in this school
16	Grade High	Highest grade level in this school
17	Community School?	Yes/No
18	School Income Estimate	Numeric
19	Percent ELL	ELL = English Language Learners
20	Percent Asian	Percentage in this ethnic group
21	Percent Black	Percentage in this ethnic group
22	Percent Hispanic	Percentage in this ethnic group
23	Percent Black / Hispanic	Percentage in this ethnic group
24	Percent White	Percentage in this ethnic group
25	Student Attendance Rate	Total number of days attended by all students / total number of days on register for all students
26	Percent of Students Chronically Absent	Missing 10% of school days - or 18 days+ per year in a 180-day school year
27	Rigorous Instruction %	How well the curriculum and instruction engage students, build critical-thinking skills, and are aligned to the Common Core
28	Rigorous Instruction Rating	Ordinal: How well the curriculum and instruction engage students, build critical-thinking skills, and are aligned to the Common Core
29	Collaborative Teachers %	How well teachers participate in opportunities to develop, grow, and contribute to the continuous improvement of the school community
30	Collaborative Teachers Rating	Ordinal: How well teachers participate in opportunities to develop, grow, and contribute to

Column	Name	Description
		the continuous improvement of the school community
31	Supportive Environment %	How well the school establishes a culture where students feel safe, challenged to grow, and supported to meet high expectations
32	Supportive Environment Rating	Ordinal: How well the school establishes a culture where students feel safe, challenged to grow, and supported to meet high expectations
33	Effective School Leadership %	How well school leadership inspires the school community with a clear instructional vision and effectively distributes leadership to realize this vision
34	Effective School Leadership Rating	Ordinal: How well school leadership inspires the school community with a clear instructional vision and effectively distributes leadership to realize this vision
35	Strong Family-Community Ties %	How well the school forms effective partnerships with families to improve the school
36	Strong Family-Community Ties Rating	How well the school forms effective partnerships with families to improve the school
37	Trust %	Whether the relationships between administrators, educators, students, and families are based on trust and respect
38	Trust Rating	Whether the relationships between administrators, educators, students, and families are based on trust and respect
39	Student Achievement Rating	Weighted Average Score + the Closing the Achievement Gap Additional Points - 4: Exceeding Target, 3: Meeting Target, 2: Approaching Target, 1: Not Meeting Target
40	Average ELA Proficiency	ELA = English Language Arts Performance Levels = 1 (insufficient), 2 (partial but insufficient), 3 (sufficient), and 4 (more than sufficient) - reflect the extent to which students demonstrate the level of understanding expected at their grade level, based on the New York State P-12 Common Core Learning Standards
41	Average Math Proficiency	
42	Grade 3 ELA - All Students Tested	No. of students tested for ELA in 3rd grade
43	Grade 3 ELA 4s - All Students	No. of students in 3rd grade who scored a 4 in ELA
44	Grade 3 ELA 4s - American Indian or Alaska Native	No. of students in 3rd grade with this particular background who scored a 4 in ELA
45	Grade 3 ELA 4s - Black or African American	No. of students in 3rd grade with this particular background who scored a 4 in ELA
46	Grade 3 ELA 4s - Hispanic or Latino	No. of students in 3rd grade with this particular background who scored a 4 in ELA
47	Grade 3 ELA 4s - Asian or Pacific Islander	No. of students in 3rd grade with this particular background who scored a 4 in ELA
48	Grade 3 ELA 4s - White	No. of students in 3rd grade with this particular background who scored a 4 in ELA

Column	Name	Description
49	Grade 3 ELA 4s - Multiracial	No. of students in 3rd grade with this particular background who scored a 4 in ELA
50	Grade 3 ELA 4s - Limited English Proficient	No. of students in 3rd grade with this particular background who scored a 4 in ELA
51	Grade 3 ELA 4s - Economically Disadvantaged	No. of students in 3rd grade with this particular background who scored a 4 in ELA
52	Grade 3 Math - All Students tested	No. of students in 3rd grade tested for Math
53	Grade 3 Math 4s - All Students	No. of students in 3rd grade who scored a 4 in math
54	Grade 3 Math 4s - American Indian or Alaska Native	No. of students in 3rd grade with this particular background who scored a 4 in math
55	Grade 3 Math 4s - Black or African American	No. of students in 3rd grade with this particular background who scored a 4 in math
56	Grade 3 Math 4s - Hispanic or Latino	No. of students in 3rd grade with this particular background who scored a 4 in math
57	Grade 3 Math 4s - Asian or Pacific Islander	No. of students in 3rd grade with this particular background who scored a 4 in math
58	Grade 3 Math 4s - White	No. of students in 3rd grade with this particular background who scored a 4 in math
59	Grade 3 Math 4s - Multiracial	No. of students in 3rd grade with this particular background who scored a 4 in math
60	Grade 3 Math 4s - Limited English Proficient	No. of students in 3rd grade with this particular background who scored a 4 in math
61	Grade 3 Math 4s - Economically Disadvantaged	No. of students in 3rd grade with this particular background who scored a 4 in math
62	Grade 4 ELA - All Students Tested	No. of students tested for ELA in 4th grade
63	Grade 4 ELA 4s - All Students	No. of students in 4th grade who scored a 4 in ELA
64	Grade 4 ELA 4s - American Indian or Alaska Native	No. of students in 4th grade with this particular background who scored a 4 in ELA
65	Grade 4 ELA 4s - Black or African American	No. of students in 4th grade with this particular background who scored a 4 in ELA
66	Grade 4 ELA 4s - Hispanic or Latino	No. of students in 4th grade with this particular background who scored a 4 in ELA
67	Grade 4 ELA 4s - Asian or Pacific Islander	No. of students in 4th grade with this particular background who scored a 4 in ELA
68	Grade 4 ELA 4s - White	No. of students in 4th grade with this particular background who scored a 4 in ELA
69	Grade 4 ELA 4s - Multiracial	No. of students in 4th grade with this particular background who scored a 4 in ELA
70	Grade 4 ELA 4s - Limited English Proficient	No. of students in 4th grade with this particular background who scored a 4 in ELA
71	Grade 4 ELA 4s - Economically Disadvantaged	No. of students in 4th grade with this particular background who scored a 4 in ELA
72	Grade 4 Math - All Students Tested	No. of students in 4th grade tested for Math
73	Grade 4 Math 4s - All Students	No. of students in 4th grade who scored a 4 in math

Column	Name	Description
74	Grade 4 Math 4s - American Indian or Alaska Native	No. of students in 4th grade with this particular background who scored a 4 in math
75	Grade 4 Math 4s - Black or African American	No. of students in 4th grade with this particular background who scored a 4 in math
76	Grade 4 Math 4s - Hispanic or Latino	No. of students in 4th grade with this particular background who scored a 4 in math
77	Grade 4 Math 4s - Asian or Pacific Islander	No. of students in 4th grade with this particular background who scored a 4 in math
78	Grade 4 Math 4s - White	No. of students in 4th grade with this particular background who scored a 4 in math
79	Grade 4 Math 4s - Multiracial	No. of students in 4th grade with this particular background who scored a 4 in math
80	Grade 4 Math 4s - Limited English Proficient	No. of students in 4th grade with this particular background who scored a 4 in math
81	Grade 4 Math 4s - Economically Disadvantaged	No. of students in 4th grade with this particular background who scored a 4 in math
82	Grade 5 ELA - All Students Tested	No. of students tested for ELA in 5th grade
83	Grade 5 ELA 4s - All Students	No. of students in 5th grade who scored a 4 in ELA
84	Grade 5 ELA 4s - American Indian or Alaska Native	No. of students in 5th grade with this particular background who scored a 4 in ELA
85	Grade 5 ELA 4s - Black or African American	No. of students in 5th grade with this particular background who scored a 4 in ELA
86	Grade 5 ELA 4s - Hispanic or Latino	No. of students in 5th grade with this particular background who scored a 4 in ELA
87	Grade 5 ELA 4s - Asian or Pacific Islander	No. of students in 5th grade with this particular background who scored a 4 in ELA
88	Grade 5 ELA 4s - White	No. of students in 5th grade with this particular background who scored a 4 in ELA
89	Grade 5 ELA 4s - Multiracial	No. of students in 5th grade with this particular background who scored a 4 in ELA
90	Grade 5 ELA 4s - Limited English Proficient	No. of students in 5th grade with this particular background who scored a 4 in ELA
91	Grade 5 ELA 4s - Economically Disadvantaged	No. of students in 5th grade with this particular background who scored a 4 in ELA
92	Grade 5 Math - All Students Tested	No. of students in 5th grade tested for Math
93	Grade 5 Math 4s - All Students	No. of students in 5th grade who scored a 4 in math
94	Grade 5 Math 4s - American Indian or Alaska Native	No. of students in 5th grade with this particular background who scored a 4 in math
95	Grade 5 Math 4s - Black or African American	No. of students in 5th grade with this particular background who scored a 4 in math
96	Grade 5 Math 4s - Hispanic or Latino	No. of students in 5th grade with this particular background who scored a 4 in math
97	Grade 5 Math 4s - Asian or Pacific Islander	No. of students in 5th grade with this particular background who scored a 4 in math
98	Grade 5 Math 4s - White	No. of students in 5th grade with this particular background who scored a 4 in math

Column	Name	Description
99	Grade 5 Math 4s - Multiracial	No. of students in 5th grade with this particular background who scored a 4 in math
100	Grade 5 Math 4s - Limited English Proficient	No. of students in 5th grade with this particular background who scored a 4 in math
101	Grade 5 Math 4s - Economically Disadvantaged	No. of students in 5th grade with this particular background who scored a 4 in math
102	Grade 6 ELA - All Students Tested	No. of students tested for ELA in 6th grade
103	Grade 6 ELA 4s - All Students	No. of students in 6th grade who scored a 4 in ELA
104	Grade 6 ELA 4s - American Indian or Alaska Native	No. of students in 6th grade with this particular background who scored a 4 in ELA
105	Grade 6 ELA 4s - Black or African American	No. of students in 6th grade with this particular background who scored a 4 in ELA
106	Grade 6 ELA 4s - Hispanic or Latino	No. of students in 6th grade with this particular background who scored a 4 in ELA
107	Grade 6 ELA 4s - Asian or Pacific Islander	No. of students in 6th grade with this particular background who scored a 4 in ELA
108	Grade 6 ELA 4s - White	No. of students in 6th grade with this particular background who scored a 4 in ELA
109	Grade 6 ELA 4s - Multiracial	No. of students in 6th grade with this particular background who scored a 4 in ELA
110	Grade 6 ELA 4s - Limited English Proficient	No. of students in 6th grade with this particular background who scored a 4 in ELA
111	Grade 6 ELA 4s - Economically Disadvantaged	No. of students in 6th grade with this particular background who scored a 4 in ELA
112	Grade 6 Math - All Students Tested	No. of students in 6th grade tested for Math
113	Grade 6 Math 4s - All Students	No. of students in 6th grade who scored a 4 in math
114	Grade 6 Math 4s - American Indian or Alaska Native	No. of students in 6th grade with this particular background who scored a 4 in math
115	Grade 6 Math 4s - Black or African American	No. of students in 6th grade with this particular background who scored a 4 in math
116	Grade 6 Math 4s - Hispanic or Latino	No. of students in 6th grade with this particular background who scored a 4 in math
117	Grade 6 Math 4s - Asian or Pacific Islander	No. of students in 6th grade with this particular background who scored a 4 in math
118	Grade 6 Math 4s - White	No. of students in 6th grade with this particular background who scored a 4 in math
119	Grade 6 Math 4s - Multiracial	No. of students in 6th grade with this particular background who scored a 4 in math
120	Grade 6 Math 4s - Limited English Proficient	No. of students in 6th grade with this particular background who scored a 4 in math
121	Grade 6 Math 4s - Economically Disadvantaged	No. of students in 6th grade with this particular background who scored a 4 in math
122	Grade 7 ELA - All Students Tested	No. of students tested for ELA in 7th grade
123	Grade 7 ELA 4s - All Students	No. of students in 7th grade who scored a 4 in ELA
124	Grade 7 ELA 4s - American Indian or Alaska Native	No. of students in 7th grade with this particular background who scored a 4 in ELA

Column	Name	Description
125	Grade 7 ELA 4s - Black or African American	No. of students in 7th grade with this particular background who scored a 4 in ELA
126	Grade 7 ELA 4s - Hispanic or Latino	No. of students in 7th grade with this particular background who scored a 4 in ELA
127	Grade 7 ELA 4s - Asian or Pacific Islander	No. of students in 7th grade with this particular background who scored a 4 in ELA
128	Grade 7 ELA 4s - White	No. of students in 7th grade with this particular background who scored a 4 in ELA
129	Grade 7 ELA 4s - Multiracial	No. of students in 7th grade with this particular background who scored a 4 in ELA
130	Grade 7 ELA 4s - Limited English Proficient	No. of students in 7th grade with this particular background who scored a 4 in ELA
131	Grade 7 ELA 4s - Economically Disadvantaged	No. of students in 7th grade with this particular background who scored a 4 in ELA
132	Grade 7 Math - All Students Tested	No. of students in 7th grade tested for Math
133	Grade 7 Math 4s - All Students	No. of students in 7th grade who scored a 4 in math
134	Grade 7 Math 4s - American Indian or Alaska Native	No. of students in 7th grade with this particular background who scored a 4 in math
135	Grade 7 Math 4s - Black or African American	No. of students in 7th grade with this particular background who scored a 4 in math
136	Grade 7 Math 4s - Hispanic or Latino	No. of students in 7th grade with this particular background who scored a 4 in math
137	Grade 7 Math 4s - Asian or Pacific Islander	No. of students in 7th grade with this particular background who scored a 4 in math
138	Grade 7 Math 4s - White	No. of students in 7th grade with this particular background who scored a 4 in math
139	Grade 7 Math 4s - Multiracial	No. of students in 7th grade with this particular background who scored a 4 in math
140	Grade 7 Math 4s - Limited English Proficient	No. of students in 7th grade with this particular background who scored a 4 in math
141	Grade 7 Math 4s - Economically Disadvantaged	No. of students in 7th grade with this particular background who scored a 4 in math
142	Grade 8 ELA - All Students Tested	No. of students tested for ELA in 8th grade
143	Grade 8 ELA 4s - All Students	No. of students in 8th grade who scored a 4 in ELA
144	Grade 8 ELA 4s - American Indian or Alaska Native	No. of students in 8th grade with this particular background who scored a 4 in ELA
145	Grade 8 ELA 4s - Black or African American	No. of students in 8th grade with this particular background who scored a 4 in ELA
146	Grade 8 ELA 4s - Hispanic or Latino	No. of students in 8th grade with this particular background who scored a 4 in ELA
147	Grade 8 ELA 4s - Asian or Pacific Islander	No. of students in 8th grade with this particular background who scored a 4 in ELA
148	Grade 8 ELA 4s - White	No. of students in 8th grade with this particular background who scored a 4 in ELA
149	Grade 8 ELA 4s - Multiracial	No. of students in 8th grade with this particular background who scored a 4 in ELA

Column	Name	Description
150	Grade 8 ELA 4s - Limited English Proficient	No. of students in 8th grade with this particular background who scored a 4 in ELA
151	Grade 8 ELA 4s - Economically Disadvantaged	No. of students in 8th grade with this particular background who scored a 4 in ELA
152	Grade 8 Math - All Students Tested	No. of students in 8th grade tested for Math
153	Grade 8 Math 4s - All Students	No. of students in 8th grade who scored a 4 in math
154	Grade 8 Math 4s - American Indian or Alaska Native	No. of students in 8th grade with this particular background who scored a 4 in math
155	Grade 8 Math 4s - Black or African American	No. of students in 8th grade with this particular background who scored a 4 in math
156	Grade 8 Math 4s - Hispanic or Latino	No. of students in 8th grade with this particular background who scored a 4 in math
157	Grade 8 Math 4s - Asian or Pacific Islander	No. of students in 8th grade with this particular background who scored a 4 in math
158	Grade 8 Math 4s - White	No. of students in 8th grade with this particular background who scored a 4 in math
159	Grade 8 Math 4s - Multiracial	No. of students in 8th grade with this particular background who scored a 4 in math
160	Grade 8 Math 4s - Limited English Proficient	No. of students in 8th grade with this particular background who scored a 4 in math
161	Grade 8 Math 4s - Economically Disadvantaged	No. of students in 8th grade with this particular background who scored a 4 in math