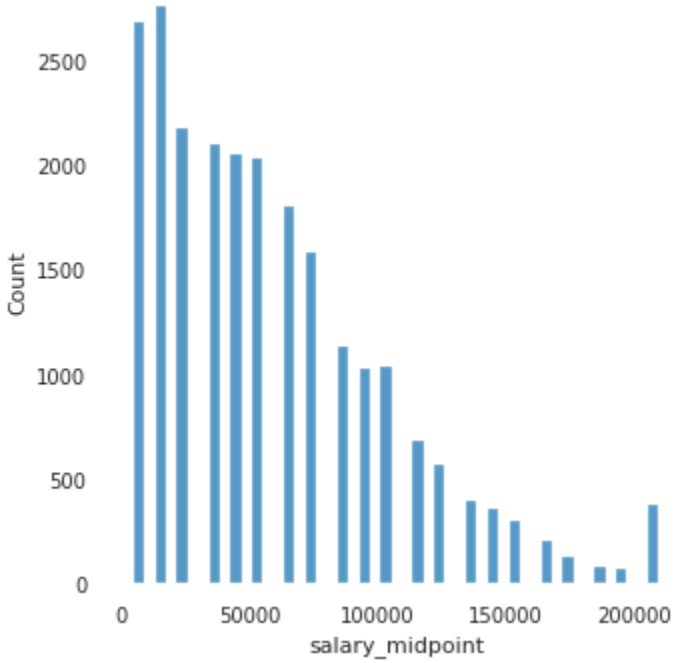
1/ Introduction and about the data

In this notebook, we will label a data scientist's salary based on information like location, gender, experience, etc. The original target value is the salary midpoint as shown in the below graph:



To simplify, we will narrow it down to groups: less than 20.000, from 20.000 to 85.000 and over 85.000 which are equivalent to label 1, 2, 3 respectively.

2/ Data Cleansing

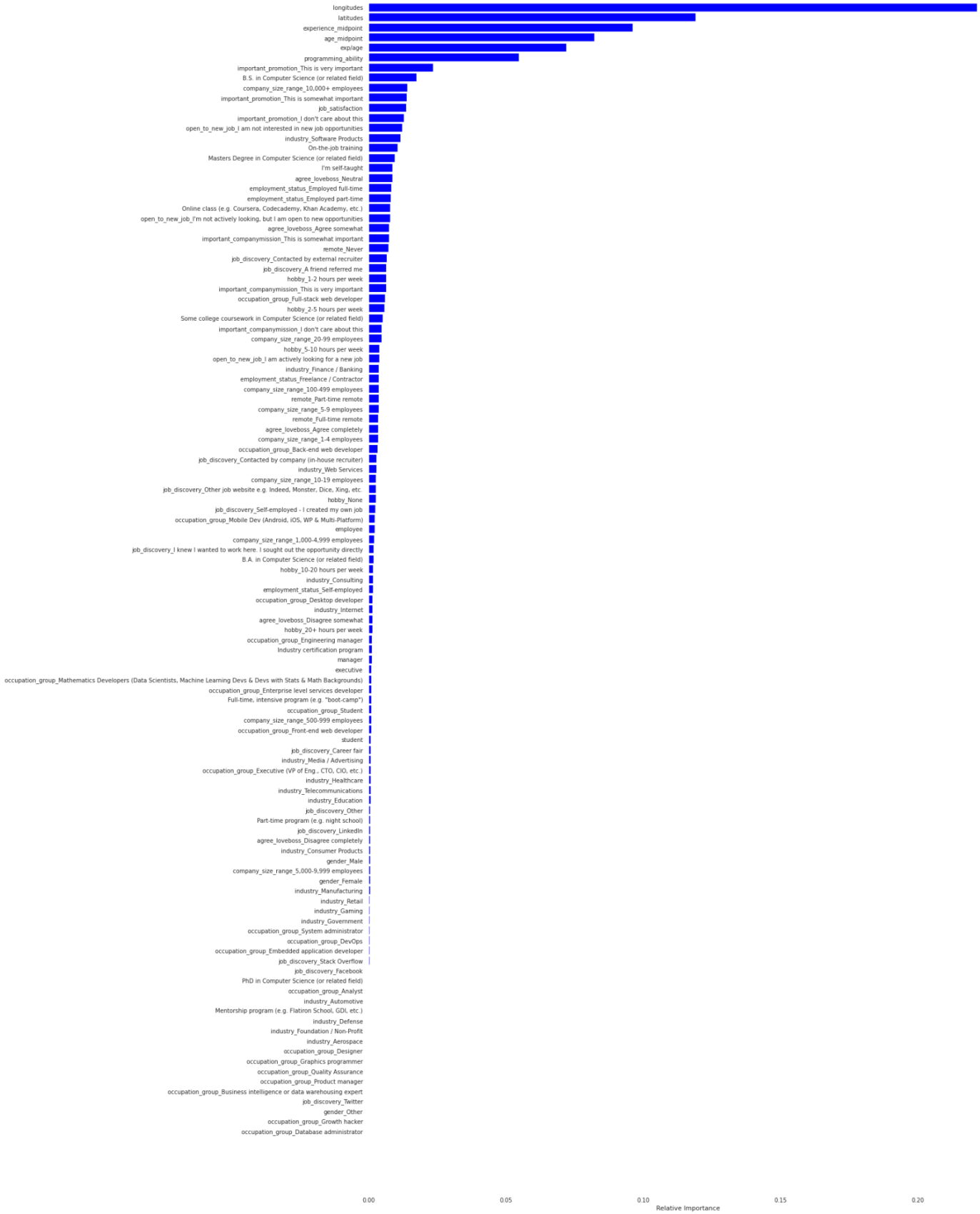
2a/ We'll first remove absolutely irrelevant features like "aliens", "dogs_vs_cats".
2b/ Then convert the country names to its coordinates.
2c/ And convert the resting features like education, gender to binary values.

3/ Splitting and Sampling

After splitting the data, you will see that we don't have much data on the first and the third group. To address this imbalance, we'll use oversampling (or undersampling, if you try both, it will be pretty much the same).

4/ Randomly choose an algorithm and Feature Selection

So far we've had 113 features in total and we want to know how important each is. I am really sorry for the chart below, it is too small to read.
My idea is that we won't manually select features and have a check at each selection. Instead, we just need to drop the last features with random ratios. For example, in the notebook, I cut off 1/4 of all features


longitudes
latitudes
experience_midpoint
age_midpoint
exp/age
programming_ability
important_promotion_This is very important
B.S. in Computer Science (or related field)
company_size_range_10,000+ employees
important_promotion_This is somewhat important
job_satisfaction
important_promotion_I don't care about this
open_to_new_job_I am not interested in new job opportunities
industry_Software Products
On-the-job training
Masters Degree in Computer Science (or related field)
I'm self-taught
agree_loveboss_Neutral
employment_status_Employed full-time
employment_status_Employed part-time
Online class (e.g. Coursera, Codecademy, Khan Academy, etc.)
open_to_new_job_I'm not actively looking, but I am open to new opportunities
agree_loveboss_Agree somewhat
important_companymission_This is somewhat important
remote_Never
job_discovery_Contacted by external recruiter
job_discovery_A friend referred me
hobby_1-2 hours per week
important_companymission_This is very important
occupation_group_Full-stack web developer
hobby_2-5 hours per week
Some college coursework in Computer Science (or related field)
important_companymission_I don't care about this
company_size_range_20-99 employees
hobby_5-10 hours per week
open_to_new_job_I am actively looking for a new job
industry_Finance / Banking
employment_status_Freelance / Contractor
company_size_range_100-499 employees
remote_Part-time remote
company_size_range_5-9 employees
remote_Full-time remote
agree_loveboss_Agree completely
company_size_range_1-4 employees
occupation_group_Back-end web developer
job_discovery_Contacted by company (in-house recruiter)
industry_Web Services
company_size_range_10-19 employees
job_discovery_Other job website e.g. Indeed, Monster, Dice, Xing, etc.
hobby_None
job_discovery_Self-employed - I created my own job
occupation_group_Mobile Dev (Android, iOS, WP & Multi-Platform)
employee
company_size_range_1,000-4,999 employees
job_discovery_I knew I wanted to work here. I sought out the opportunity directly
B.A. in Computer Science (or related field)
hobby_10-20 hours per week
industry_Consulting
employment_status_Self-employed
occupation_group_Desktop developer
industry_Internet
agree_loveboss_Disagree somewhat
hobby_20+ hours per week
occupation_group_Engineering manager
Industry certification program
manager
executive
occupation_group_Mathematics Developers (Data Scientists, Machine Learning Devs & Devs with Stats & Math Backgrounds)
occupation_group_Enterprise level services developer
Full-time, intensive program (e.g. "boot-camp")
occupation_group_Student
company_size_range_500-999 employees
occupation_group_Front-end web developer
student
job_discovery_Career fair
industry_Media / Advertising
occupation_group_Executive (VP of Eng., CTO, CIO, etc.)
industry_Healthcare
industry_Telecommunications
industry_Education
job_discovery_Other
Part-time program (e.g. night school)
job_discovery_LinkedIn
agree_loveboss_Disagree completely
industry_Consumer Products
gender_Male
company_size_range_5,000-9,999 employees
gender_Female
industry_Manufacturing
industry_Retail
industry_Gaming
industry_Government
occupation_group_System administrator
occupation_group_DevOps
occupation_group_Embedded application developer
job_discovery_Stack Overflow
job_discovery_Facebook
PhD in Computer Science (or related field)
occupation_group_Analyst
industry_Automotive
Mentorship program (e.g. Flatiron School, GDI, etc.)
industry_Defense
industry_Foundation / Non-Profit
industry_Aerospace
occupation_group_Designer
occupation_group_Graphics programmer
occupation_group_Quality Assurance
occupation_group_Product manager
occupation_group_Business intelligence or data warehousing expert
job_discovery_Twitter
gender_Other
occupation_group_Growth hacker
occupation_group_Database administrator
0.00
0.05
0.10
0.15
0.20
Relative Importance

And here is the results comparison:

	set	label 1	label 2	label 3	f1_score
0	train_set	0.87	0.78	0.78	0.86
1	val_set	0.86	0.78	0.78	0.77
2	test_set	0.85	0.74	0.74	0.77

Using all features

	set	label 1	label 2	label 3	f1_score
0	train_set	0.8	0.78	0.79	0.8
1	val_set	0.8	0.77	0.78	0.77
2	test_set	0.78	0.74	0.76	0.78

Using 3/4 features (This is a bit more reasonable)

This result is acceptable in general. But as you can see, there is no difference between dropping or not dropping 1/3 of the features. Please let me know if you know why it is. I also tried dropping rates of 1/2 and 1/4 but the performance was worse.