Business Understanding

Currently, skills & competencies acqusition processes tend
to be fragmented and remain arbitary, based on expert opinions

ML solution can help our users see the potential value of specific knowledge portfolio and become more engaged and confident with learning process.

Business Understanding

ML Task Type: Regression

We predict continuous value – salary in $ for a skillset in job listing

Metric Threshold

10%

Root Mean Squared Error

To "punish" model for overly big errors. Big error → Failed user expectations → Trust loss.

MAPE

Interpretable. To assess model in the business sense. On average, how far are our predictions from the actual values?

Project Plan

Business understanding

Get the context of the problem and examine user jtbd

Select ML approach that fits the business task

Select an ML model metric in accordance with business goal and context

Specify the desired metric threshold out of business need

Define Roles & Responsibilities of project Team

Data Collection

Research external job listing sources

Assess gathered sources by criteria:

- Website traffic
- Sample size
- Availability of salary (fork) for vacancies
- Specialization scope
- Presence of skills structuring / tags
- Geo coverage
- Presence of vacancy grades
- Perceived "ease" of scraping

Choose source(s) from pool by criteria

Scrape data from external source(s)

Transfer scraped data to data structure

Data Preprocessing

Extract salary from textual job description by a string pattern (for listings without salary "tag")

Perform data cleaning:

- Translate to English if any
- Clear duplicates
- Check for NaNs

Perform EDA to improve data understanding

Preprocess textual job description:

- Clear it from company name
- Clear it from salary mentions
- Fill in a dictionary of stop words relevant to the task
- Filter out stop words

Convert string salary to numeric value

Encode categorical features:

- skills/technologies tags
- position grade tags
- job category tags

Modelling

Extract features from textual job description using embedding:

- word2vec
- GloVe

Choose train-test-validation split approach

Build naive model as baseline for comparison

Research modeling approaches for similar cases

Develop competitive regression models:

- CatBoost Regressor
- Linear Regressor
- k-nearest neighbours Regressor
- Bayesian Regressor
- to be continued

Evaluation

Compare models' results by metric(s)

Select best performing model

Run Grid Search to find best set of parameters for selected model

Business Value

Evaluate business effect:

Build financial model and evaluate costs

Devise deployment strategy

Data Collection

Data Source:

Chosen ai-jobs.net for our purposes

1. Salary fork
2. International
3. Tech specific
4. Skills/tools/requirements/stack tags
5. Relatively easy to scrape

Libraries used:
- Selenium
- BeautifulSoup

Issues:
- Exceptions
- Processing in chunks

Scraped:
3754 job postings

Step 1:

scraped index page for all available job listing URLs with key descriptions (position, salary, company)

	url	title	location	type	level	salaryRange	salary	company
0	https://ai-jobs.net/job/32218-data-science-con...	View full details of `Data Science Content Int...	Remote	Internship	Entry-level	USD 9K - 11K	NaN	NannyML
1	https://ai-jobs.net/job/32029-python-php-senio...	View full details of `Python / PHP Senior Deve...	Remote, EU	Full Time	Senior-level	EUR 65K - 80K	NaN	Beatopia
2	https://ai-jobs.net/job/30723-data-scientist-r...	View full details of `Data Scientist - Researc...	Remote, Hybrid Available (Chicago)	Full Time	Senior-level	USD 55K - 120K	NaN	HFR, Inc.

Step 2:

from the list of URLs gathered at previous step went one by one, scraping each posting's details:
- Minimum and Maximum Salary
- Position
- Date of Posting and Expiration Date
- Job Descriptions
- Skills from Tags

	minSalary	maxSalary	currency	salaryPeriod	position	datePosted	validThrough	jobDescription	skills
0	230400.0	345600.0	USD	YEAR	Director, Engineering Data Architecture	2022-10-31 16:35:17	2022-12-15 00:00:00	Why Glassdoor? Our mission is to help people e...	[Agile, Airflow, AWS, Big Data, Computer Scien...
1	NaN	NaN	None	None	Data Operations Analyst II (R-13164)	2022-10-31 15:58:11	2022-12-15 00:00:00	Why We Work at Dun & BradstreetDun &am...	[Agile, AWS, Computer Science, Data analysis, ...
2	42000.0	78000.0	USD	YEAR	Data Engineer (Remote)	2022-10-31 15:56:06	2022-12-15 00:00:00	This is a full-time, fully remote role for Lat...	[Airflow, AWS, Big Data, Bitbucket, Computer S...

Data Preprocessing

What	How
Extracted Currency from string salary fork	
Extracted numeric salary values from string salary fork	
Extracted numeric salary (for some only minimum, for some - whole fork) from text job description	pandas.Series.str.findall (pattern)
Converted salaries in other currencies (EUR, GBP) to USD	
Detected language for text job descriptions	pycld2 for language detection
Translated text job descriptions to English from other languages (through tokenization into sentences, sentence by sentence)	GoogleTranslator from deep_translator
Encoded skills from tags into binary variables	pandas.DataFrame.explode + pandas.crosstab
Clear out emojis and other special symbols from text job description	re
Cleaned and merged Position and Title	
Removed instances without Salary and Job Description	

Sample size

2006 rows with salary – target label

Data Preprocessing – Text

Cleared:

- Emojis
- Pictograms
- Special Unicode characters from scraping
- Punctuation
- Stopwords

Lowercased
Lemmatized

Libraries used:

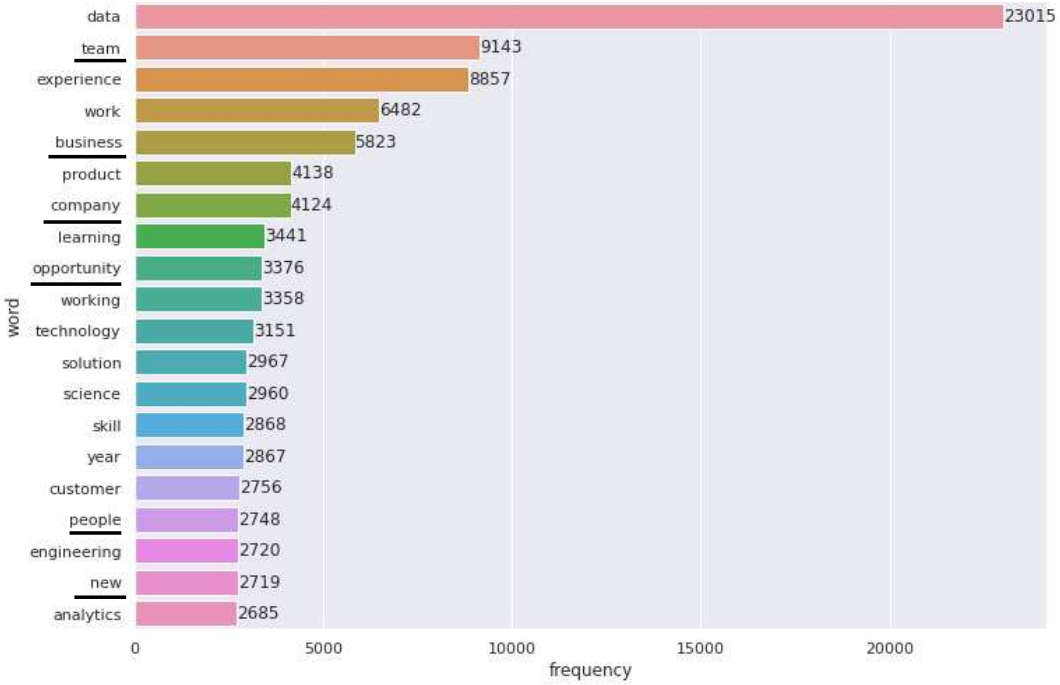
- re
- Html
- WordNetLemmatizer from nltk.stem

'Description NannyML - creators of an Open Source Python library, are looking for multiple Data Science Write r interns to help across Data Science content creation, research and prototyping. About Us NannyML is an Open Source Python library for detecting silent model failure. We are backed by tech leaders who have founded and grown top data unicorns, like the CTO of Collibra (the first Belgian unicorn), Zehan Wang (ex-head of ML at T witter); and deep tech VCs like Lunar Ventures. At NannyML, we are building Free Open Source Software to esti mate the performance of ML models after deployment. We're creating a one-of-a-kind company by being one of th e initial players in the growing ML Monitoring field. By focusing on post-deployment data science and perform ance estimation, we'll define the industry and science for decades to come. The founders are experienced entr epreneurs that previously founded a specialised machine learning company, where they became experts at buildi ng machine learning systems. At NannyML, we take pride in hiring the best people and getting out of their way so that they can make great things happen. \xa0 About the Role We are looking for Data Science Writer interns to help the Research and Growth team create niche-defining content. You will work with senior Data profession als and all 3 founders, as well as everyone else at the company. You will be responsible for creating new con tent from scratch, curating content across multiple sources and iterating on blog outlines created by other t eam members. As we grow NannyML we expect you to grow with us. As our company scales, we envision a successfu l candidate to grow into a Data Science Writer or a similar full-time position. Requirements ● You've created multiple long-form Data Science pieces of content (like blogs, articles, courses etc.) ○ Please submit a link to your blog/content as part of your application. ● Exceptional communication skills in English - both oral a nd written. ● Strong theoretical and practical understanding of Machine Learning, including hands-on experien ce in developing ML systems. ● Knowledge of the Python Data Science stack including pandas, and Scikit Learn and visualization tools. ● You are a swift learner and can easily pick up new concepts. ● You are incredibly proactive, independent and comfortable in proposing new ideas. This also means holding your ground when you b elieve you are right. ● Be available to start in the next three weeks at most. We're looking to fill thi s role immediately. ● You live in or are willing and able to move to EU time zones, and you are open to trave l roughly once per quarter Nice to haves ● STEM background ● Prior work or internship experience in a Machine Learning company ● Track record of open-source contributions Benefits ● The opportunity to be a part of the e xciting early stages of a well-funded, European-based Open Source start-up that has massive growth and ventur e potential ● Fully Remote Working Environment ● 50€/month development budget to learn Data Science, Causal M L, Bayesian Inference or anything you like that applies to your role. ● 45€/month well-being allowance (for y oga, gym etc.) ● Compensation: up to 750 EUR/month for half-time ♥ 5 day long company off-site once per quart er (all inclusive) \xa0 Our Values We value freedom with responsibility, transparency, and a growth mindset. We believe in generating our own luck by trying out new stuff, always asking, constantly learning, reading, a nd meeting new people with different world-views. We appreciate that from time to time, things may break. Wor king at NannyML, you will have full autonomy to make impactful decisions and prioritize and organize your wor k the way you see fit. You will work closely with the founders. ♥\xa0Why you will love to join NannyML ● Wor king with a fast-growing international VC funded startup with a flat structure ● You will have full ownership of the things you work on ● An international and diverse team ● Fully remote work, with as many opportunities to meet up as you want! ● Open Source Product [Link] with all the perks that come with that 😡\xa0What you m ight not love ● We're a small team, priorities will keep on changing and processes are not fine-tuned ye t ● Really fast-paced environment with a LOT of work to be done ● You will have to learn new things all the t ime \xa0 Support our open-source library by 🌟 staring us on Github'

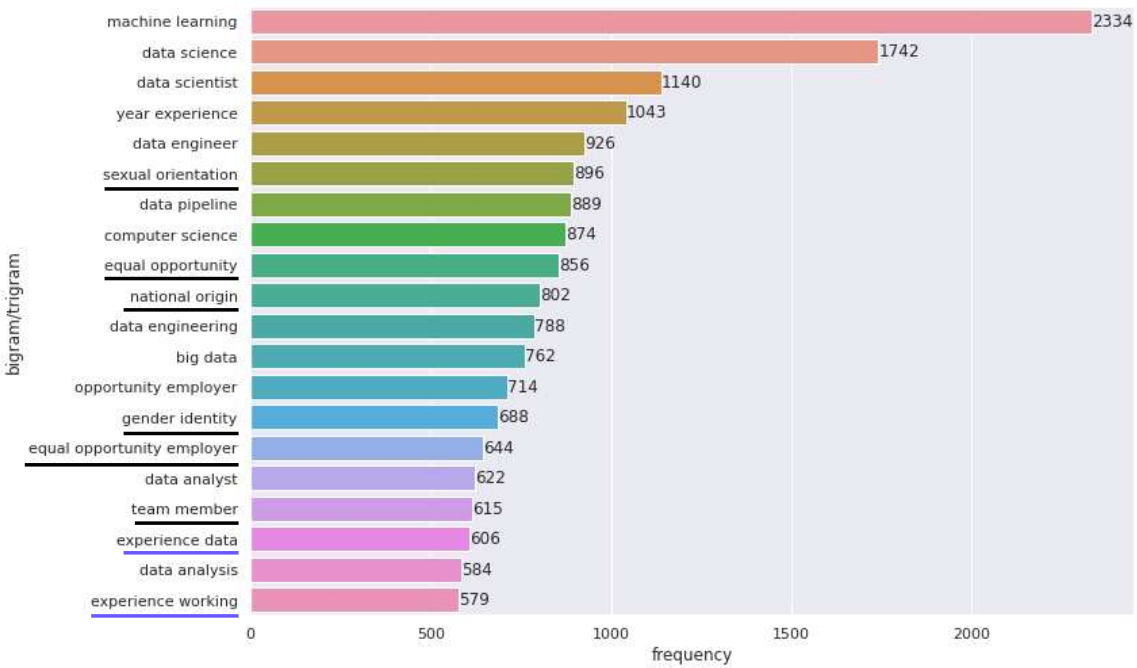
'description nannyml creator open source python library looking multiple data science writer intern help across da ta science content creation research prototyping u nannyml open source python library detecting silent model failu re backed tech leader founded grown top data unicorn like cto collibra first belgian unicorn zehan wang exhead ml twitter deep tech vcs like lunar venture nannyml building free open source software estimate performance ml model deployment creating oneofakind company one initial player growing ml monitoring field focusing postdeployment data science performance estimation well define industry science decade come founder experienced entrepreneur previousl y founded specialised machine learning company became expert building machine learning system nannyml take pride h iring best people getting way make great thing happen role looking data science writer intern help research growth team create nichedefining content work senior data professional 3 founder well everyone else company responsible c reating new content scratch curating content across multiple source iterating blog outline created team member gro w nannyml expect grow u company scale envision successful candidate grow data science writer similar fulltime posi tion requirement youve created multiple longform data science piece content like blog article course etc please su bmit link blogcontent part application exceptional communication skill english oral written strong theoretical pra ctical understanding machine learning including handson experience developing ml system knowledge python data scie nce stack including panda scikit learn visualization tool swift learner easily pick new concept incredibly proacti ve independent comfortable proposing new idea also mean holding ground believe right available start next three we ek were looking fill role immediately live willing able move eu time zone open travel roughly per quarter nice ste m background prior work internship experience machine learning company track record opensource contribution benefi t opportunity part exciting early stage wellfunded europeanbased open source startup ha massive growth venture pot ential fully remote working environment 50month development budget learn data science causal ml bayesian inference anything like applies role 45month wellbeing allowance yoga gym etc compensation 750 eurmonth halftime 5 day long company offsite per quarter inclusive value value freedom responsibility transparency growth mindset believe gener ating luck trying new stuff always asking constantly learning reading meeting new people different worldviews appr eciate time time thing may break working nannyml full autonomy make impactful decision prioritize organize work wa y see fit work closely founder love join nannyml working fastgrowing international vc funded startup flat structur e full ownership thing work international diverse team fully remote work many opportunity meet want open source pr oduct link perk come might love were small team priority keep changing process finetuned yet really fastpaced envi ronment lot work done learn new thing time support opensource library staring u github'

EDA – Text Job Descriptions

20 most popular words in Job Descriptions:



20 most popular bigrams/trigrams in Job Descriptions:



Libraries used:

- nltk – for stopwords
- sklearn - CountVectorizer

Further need to:

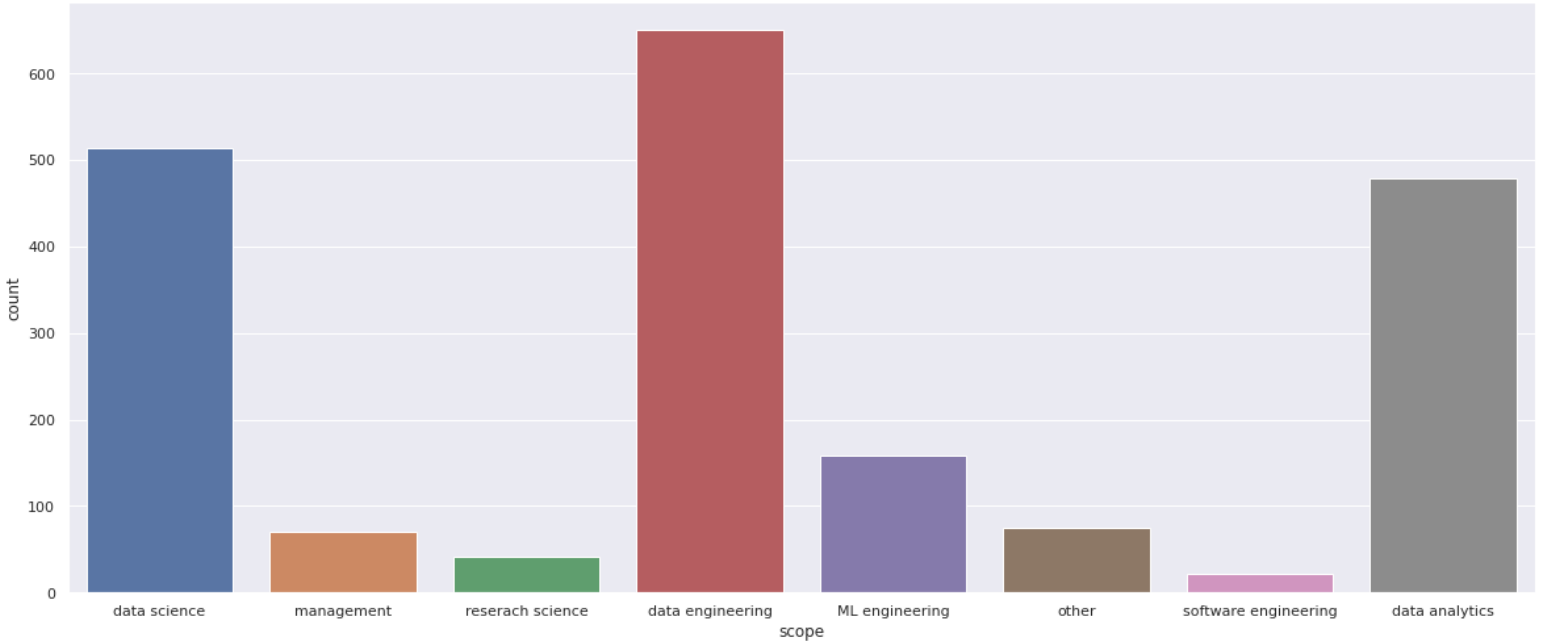
- Add task-specific generic words to stopwords list
- Experiment with extracting skills

EDA – New Scope field

After preprocessing of 'Position' field

	frequency	position
0	610	data engineer
1	492	data scientist
2	470	data analyst
3	158	ml engineer
4	64	data engineering
5	45	data science
6	43	research scientist
7	39	sr data
8	30	staff data
9	29	big data
10	27	software engineer
11	23	engineer ii
12	23	engineer data
13	21	big data engineer
14	19	product data
15	19	manager data

Constructed additional feature 'Scope':



Libraries used:

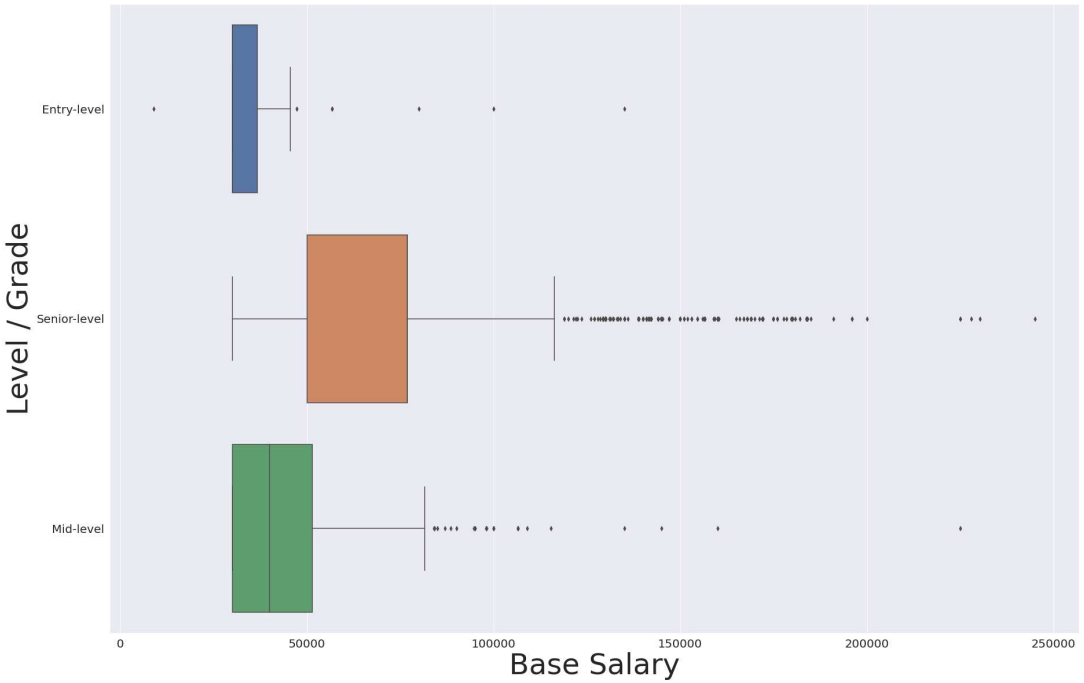
- sklearn - CountVectorizer
- seaborn

Examples of Position in 'Other':

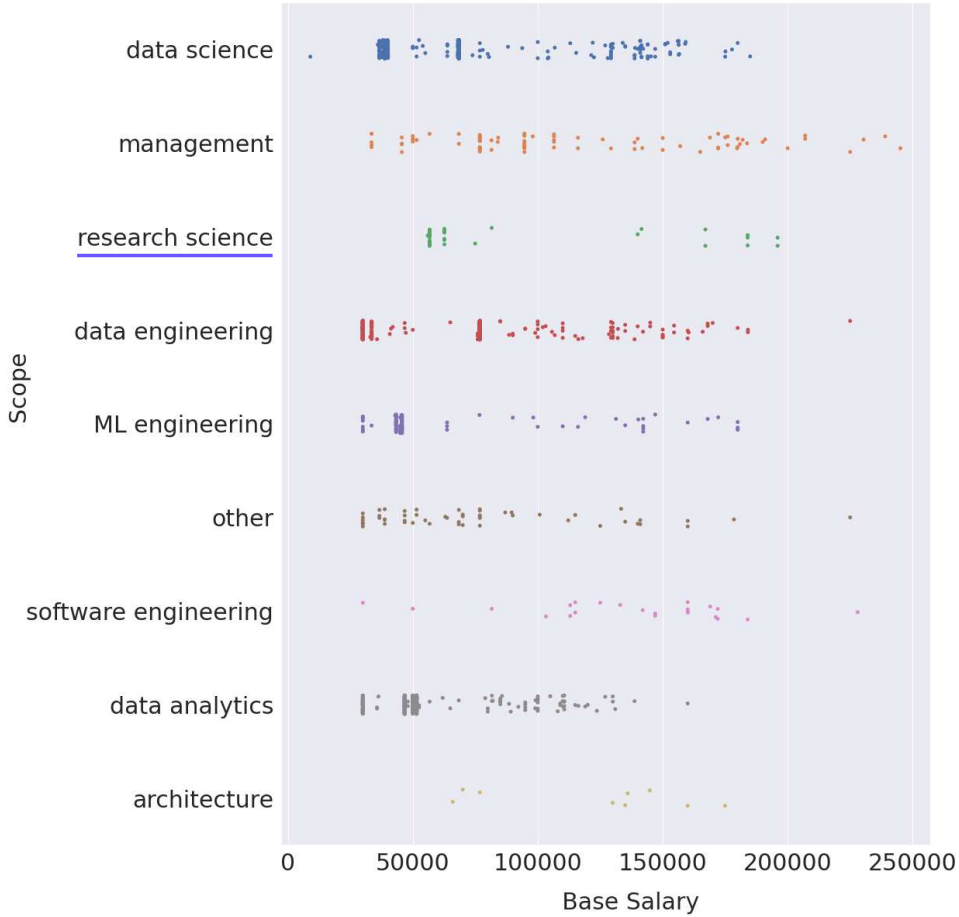
- Power BI Developer
- Head of Data Security
- Consultant

EDA – Salary (target)

Base salary by Grade: Entry / Mid / Senior



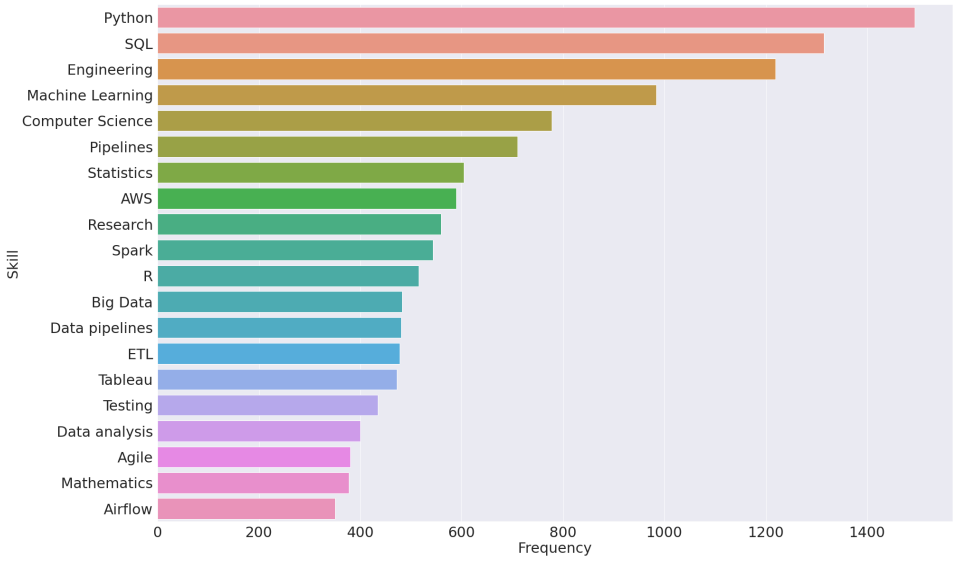
Base salary by Level: Entry / Mid / Senior



EDA – Skill Tags

Skills from tags: 284 unique skill tags

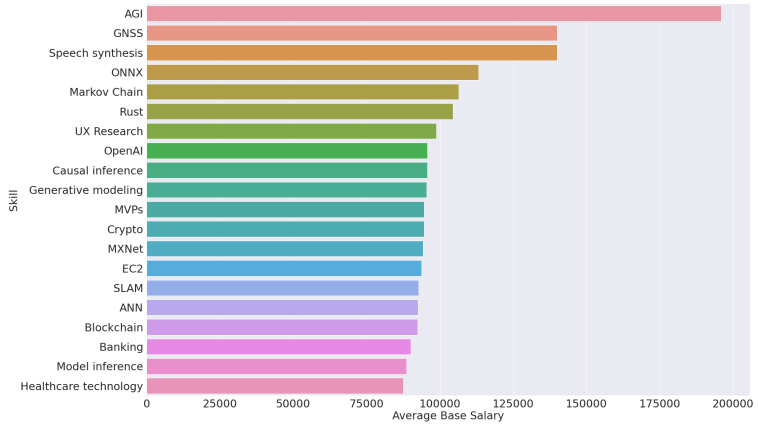
```
['Bayesian', 'Content creation', 'GitHub', 'Machine Learning', 'ML models', 'Open Source', 'Pandas', 'Prototyping'
, 'Python', 'Research', 'Scikit-learn', 'STEM']
['Computer Science', 'Data Analytics', 'Data visualization', 'Engineering', 'Hadoop', 'Machine Learning', 'MySQL',
'Predictive modeling', 'Python', 'R', 'Research', 'Spark', 'Statistics', 'TensorFlow']
['Agile', 'Airflow', 'APIs', 'AWS', 'CI/CD', 'Computer Science', 'Data management', 'Machine Learning', 'Pipelines
', 'Python', 'Redshift', 'SQL', 'STEM', 'Tableau', 'TDD', 'XGBoost']
['Data analysis', 'Economics', 'Excel', 'Research', 'SAS', 'Statistics']
```



Insight for further project task (outside the course scope):

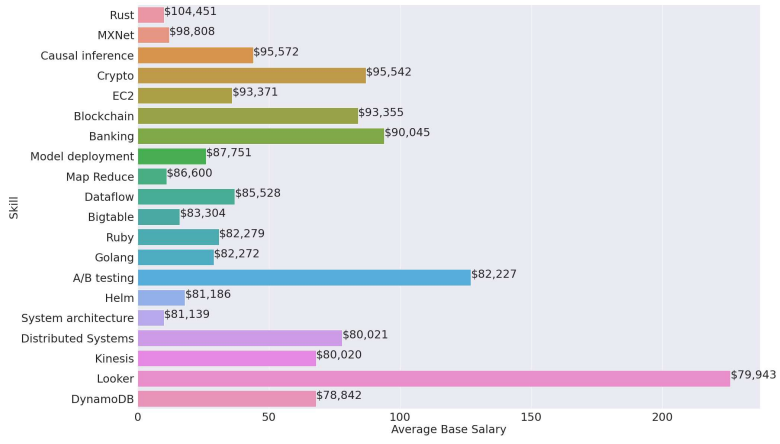
- Try to build weighed coefficient to detect most trending and simultaneously most paying skills

Average salary by skill:



For skills encountered in at least 10 job postings:

Sorted by salary (desc)



EDA – Skill Tags

20 most associated skills:

Skill1	Skill2	Pearson's R
Data pipelines	Pipelines	0.758771
TensorFlow	PyTorch	0.756263
SLAM	GNSS	0.706930
GNSS	SLAM	0.706930
Talend	Informatica	0.666415
ICLR	Human Machine Interaction	0.654000
ICLR	3D Reconstruction	0.654000
NeurIPS	ICML	0.653179
ICLR	NeurIPS	0.653179
NeurIPS	ICLR	0.653179
Blockchain	Crypto	0.651941
Google Cloud	GCP	0.622198
GCP	Google Cloud	0.622198
NumPy	Pandas	0.612825
HBase	Cassandra	0.599165
Cassandra	HBase	0.599165
Kubernetes	Docker	0.577249
Docker	Kubernetes	0.577249
Anaconda	KNIME	0.577062
Nvidia Jetson	ONNX	0.577062

20 least associated skills:

Skill1	Skill2	Pearson's R
Kafka	Mathematics	-0.150964
Keras	SQL	-0.151339
Research	ETL	-0.154507
Statistics	Kafka	-0.155491
Kafka	Statistics	-0.155491
Statistics	Pipelines	-0.156301
Research	SQL	-0.157807
Tableau	PhD	-0.158568
Data pipelines	Statistics	-0.162457
Statistics	Data pipelines	-0.162457
R	Spark	-0.162849
Mathematics	Airflow	-0.164424
Tableau	Spark	-0.166012
Data pipelines	Research	-0.166770
Research	Data pipelines	-0.166770
Engineering	Excel	-0.172388
Kafka	Research	-0.173306
Research	Kafka	-0.173306
Deep Learning	SQL	-0.210313

Most Popular Skillsets:

Skillset	Frequency	AvBaseSalary
['Engineering', 'Pipelines', 'Python', 'SQL']	367	69755
['Engineering', 'Machine Learning', 'Python', 'SQL']	356	69072
['Data pipelines', 'Pipelines', 'Python', 'SQL']	354	68018
['Data pipelines', 'Engineering', 'Pipelines', 'Python']	344	70370
['Computer Science', 'Engineering', 'Machine Learning', 'Python']	338	68273
['Computer Science', 'Engineering', 'Python', 'SQL']	337	67986
['Data pipelines', 'Engineering', 'Pipelines', 'SQL']	319	70415
['Data pipelines', 'Engineering', 'Python', 'SQL']	288	69905
['Engineering', 'Python', 'SQL', 'Spark']	280	69443
['AWS', 'Engineering', 'Python', 'SQL']	278	72418

Most Popular with Average Base Salary > q (0.9):

Skillset	Frequency	AvBaseSalary	Scope
['Data visualization', 'Looker', 'Python', 'SQL']	50	104769	data science
['Data visualization', 'Looker', 'Python', 'Tableau']	47	103663	data science
['A/B testing', 'Python', 'Statistics', 'Testing']	45	98120	data science
['Banking', 'Blockchain', 'Crypto', 'SQL']	38	122077	data science
['Banking', 'Blockchain', 'Crypto', 'Python']	37	128317	data science
['Banking', 'Blockchain', 'Crypto', 'Engineering']	37	122921	data science
['AWS', 'Computer Science', 'Engineering', 'Streaming']	36	96278	data engineering
['Banking', 'Blockchain', 'Crypto', 'Looker']	36	129803	data science
['Looker', 'Mathematics', 'Python', 'SQL']	36	98937	data science
['Blockchain', 'Crypto', 'Python', 'SQL']	35	116909	data science

Finalizing Data Preprocessing

Reducing # of Skills features:

- Cosine similarity metric calculation
- Visual assessment

Integrated:

skill1	skill2	cosine
Consulting	Consulting firm	0.414421
Data pipelines	Pipelines	0.793448
GCP	Google Cloud	0.607332
ML models	Machine Learning	0.536746

Dropped:

- PhD
- Travel

as not skills

302 skills tags → 296 skills tags

Encoding categorical features:

OneHotEncoding for:

- **Type:** Full Time / Part Time / Internship
- **Level:** Entry / Mid / Senior
- **Scope:** data science / data engineering / …

Dealing with empty Job Description:

Had **35** rows with no job description, yet we wanted to use it for prediction
Dropped them

Left with **3 719** rows, of them:
- **1 713** don't have salary mentioned
- **2 006** have salary mentioned

Modeling Strategy

Reducing # of Skills features:

- Try different approaches
- No extensive GridSearches and rigorous parameter tuning at this stage
- Split responsibility zones within the team according to features used in modeling

Selected Regressors:

- Linear Regressor (scikit-learn)
- Random Forest Regressor (scikit-learn)
- KNN Regressor (scikit-learn)
- CatBoost Regressor

Naive Baseline:

For model always predicting mean value:

RMSE = 35 256.1268
MAPE = 44.6256%

Skills as binary features:

.NET	3D Reconstruction	A/B testing	AGI	AI governance	AI strategy	AIStats	ANN	...	Unstructured data	VR	Visual SLAM	Weka	XGBoost	XML	ggplot2	spaCy
0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Categorical features:

type	level	scope
Full Time	Senior	data engineering
Full Time	Senior	data analytics
Full Time	Senior	data science
Full Time	Senior	data engineering
Full Time	Senior	data science
...	...	...

Text feature Job Description:

finalJobDescription
believe better way choose transparent le confu...
techenabled logistics mission provide hasslefr...
global blockchain behind world largest digital...
bosch global software technology private limit...
building world talent e first talent lifecycle...
...

Interim Results

Naive Baseline: 35 256.1268 44.6256%

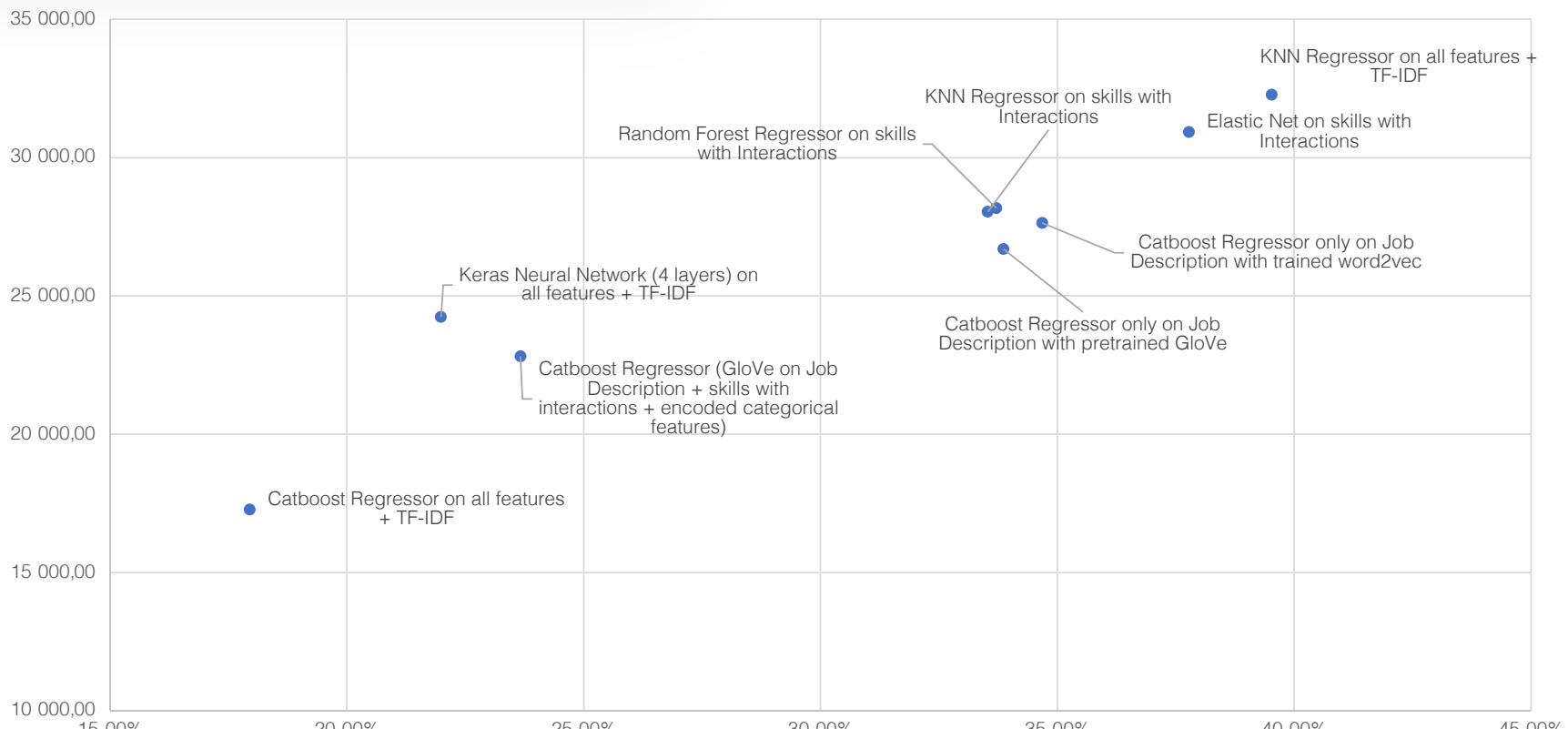
Features used in model	Regressor	Test RMSE	Test MAPE	Best params (for now)
Only Skills, transformed with Polynomial features to get pair-wise interactions	Elastic Net (sklearn)	30 919.0492	37.7793%	alpha = 0.5 l1_ratio = 0.5
Only Skills → with Polynomial features	Random Forest Regressor (sklearn)	28 174.1195	33.7082%	n_estimators = 200 max_depth = 200 min_samples_leaf = 1 max_features = 'log2'
Only Skills → with Polynomial features	KNN Regressor (sklearn)	28 039.1691	33.5332%	n_neigbours = 15 algorithm = 'ball_tree' metric = 'braycurtis'
• **Skills** → with Polynomial features **Categorical (Level, Type, Scope)** → One Hot Encoded • **Job Description (text)** → TF-IDF	KNN Regressor (sklearn)	32 268.2109	39.5282%	knn_reg__algorithm = 'brute' knn_reg__metric = 'cityblock' knn_reg__n_neighbors = 7 tfidf__max_df = 0.97
• **Skills** → with Polynomial features to get pair-wise interactions • **Categorical (Level, Type, Scope)** → One Hot Encoded • **Job Description (text)** → TF-IDF	Catboost Regressor	**17 278.1585**	**17.9491%**	learning_rate = 0.15 depth = 8 l2_leaf_reg = 1
Only Job Description (text vectorized with *pretrained* GloVe)	Catboost Regressor	26 688.3354	33.8636%	learning_rate = 0.15 depth = 8 l2_leaf_reg = 0.5
Only Job Description (text vectorized with *trained* word2vec)	Catboost Regressor	27 633.0326	34.6797%	learning_rate = 0.15 depth = 10 l2_leaf_reg = 3
• **Skills** → with Polynomial features **Categorical (Level, Type, Scope)** → One Hot Encoded • **Job Description (text)** → *pretrained* GloVe	Catboost Regressor	**22 809.7165**	**23.6633%**	learning_rate = 0.15 depth = 10 l2_leaf_reg = 0.5

Modeling

Naive Baseline:

For model always predicting mean value:

RMSE = 35 256.1268
MAPE = 44.6256%



Evaluation & Grid Searches

Naive Baseline:	35 256.1268	44.6256%
Previous Best:	17 278.1585	17.9491%

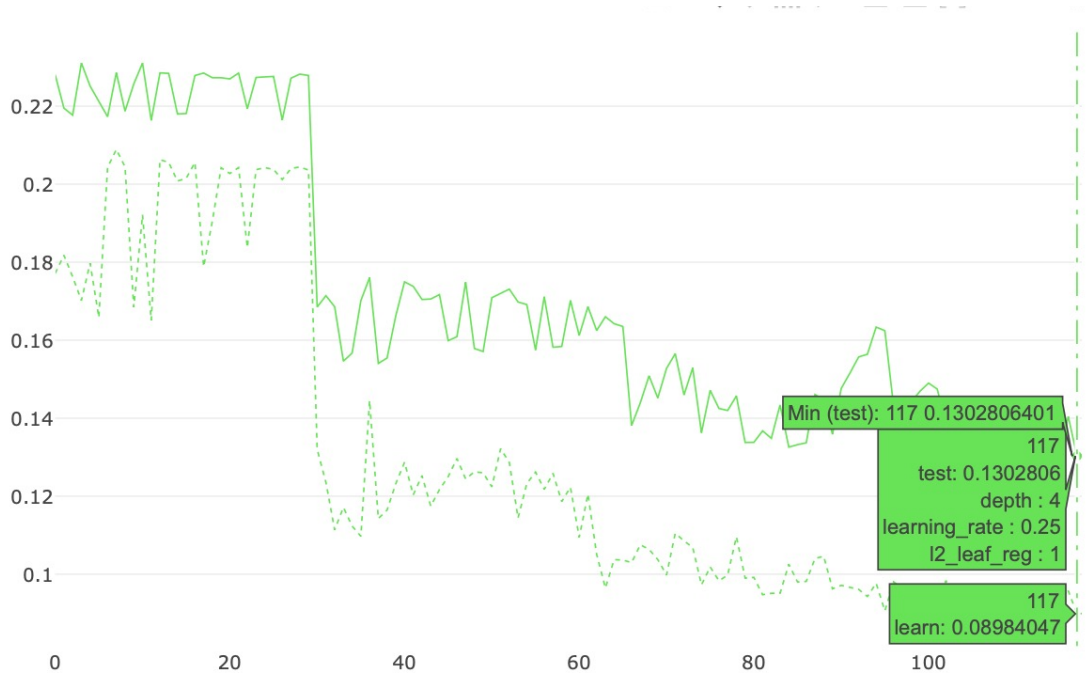
Loss function: Huber Loss (delta = 1.35)

In total, ran over these parameters in all GridSearches:
'learning_rate': [0.03, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35]
'depth': [1, 2, 3, 4, 6, 8, 10, 12]
'l2_leaf_reg': [0.1, 0.25, 0.5, 0.75, 1, 3, 5, 7, 9]

Refitted best nodel found by GridSearch:
iterations=30000
early_stopping_rounds=50

loss_function='Huber:delta=1.35'
depth=4
learning_rate=0.2
l2_leaf_reg=1

	train	test	val
rmse	19684.607243	19380.144548	17857.495810
mape	0.077916	0.122413	0.117907
r2	0.692219	0.696637	0.719642



Worst with depth = 1 (as expected)

Final model

Using previous model we filled in salary for those job postings that didn't have one

We checked the MAPE metric for initially labeled data just to make sure model works right:
MAPE= 9.1334%

We then retrained model on our new, "filled", wholly labeled dataset:

iterations=30000
early_stopping_rounds=10

loss_function='Huber:delta=1.35'
depth=4
learning_rate=0.2
l2_leaf_reg=1

	train	test	val
rmse	15085.240756	19283.976536	17216.440051
mape	0.069517	0.095631	0.116609
r2	0.711114	0.585710	0.652358