

TSDN 2022 FromScratch

Resume Reader and Salary Predictor



FromScratch Team

The Team



Danang Rizky Nugroho



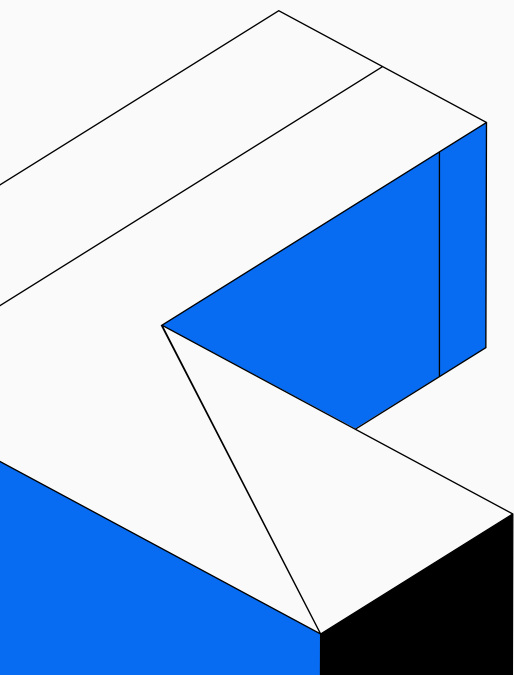
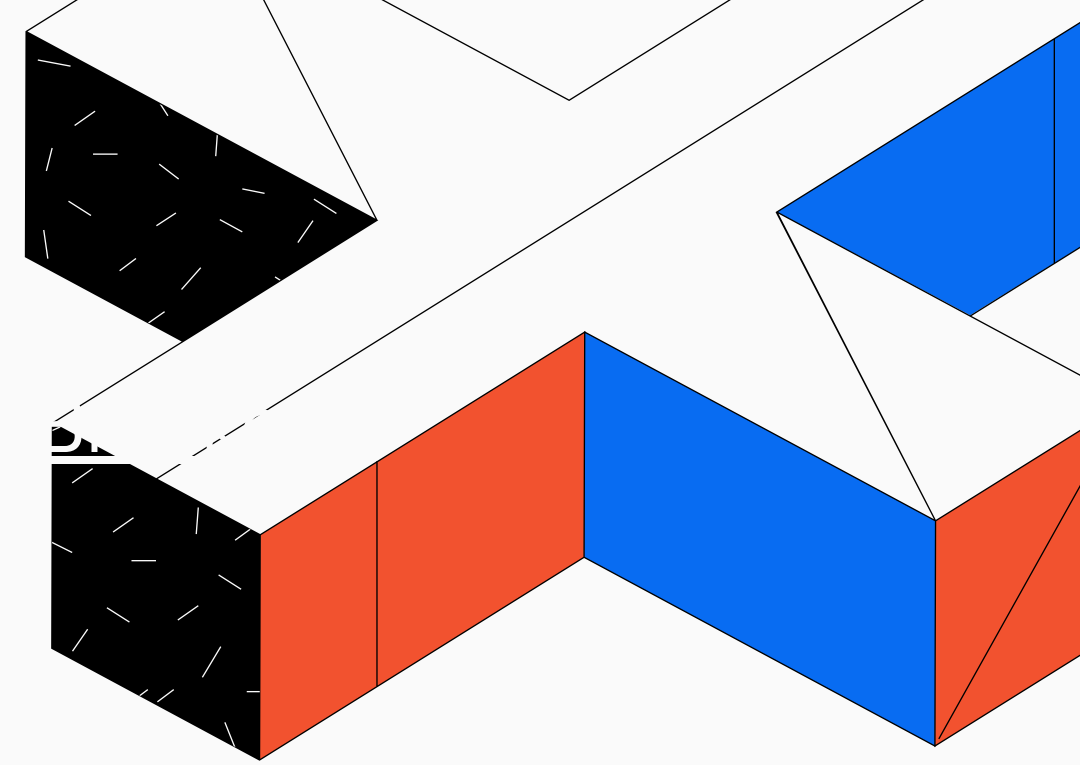
Maya Maryanah



Inggriani Priscilia



Mohammad Ifaizul Hasan

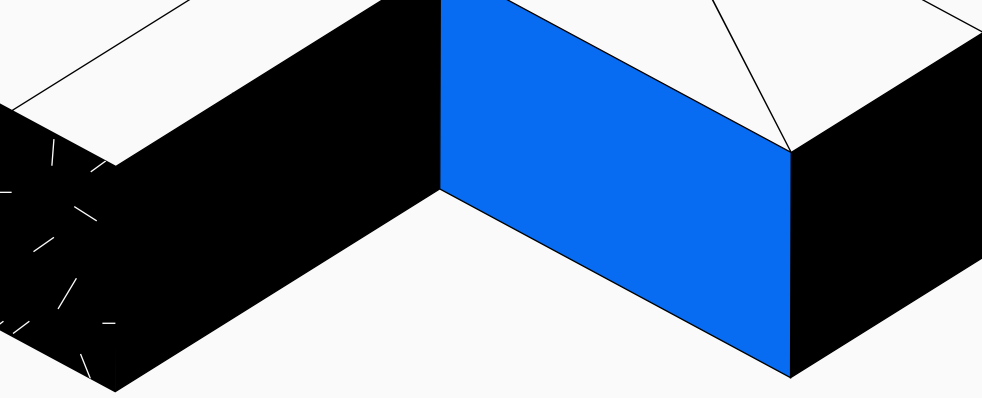




Content List



- Background
- Reason of Usecase
- About Dataset
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Modelling
- Business Recommendation
- The Products



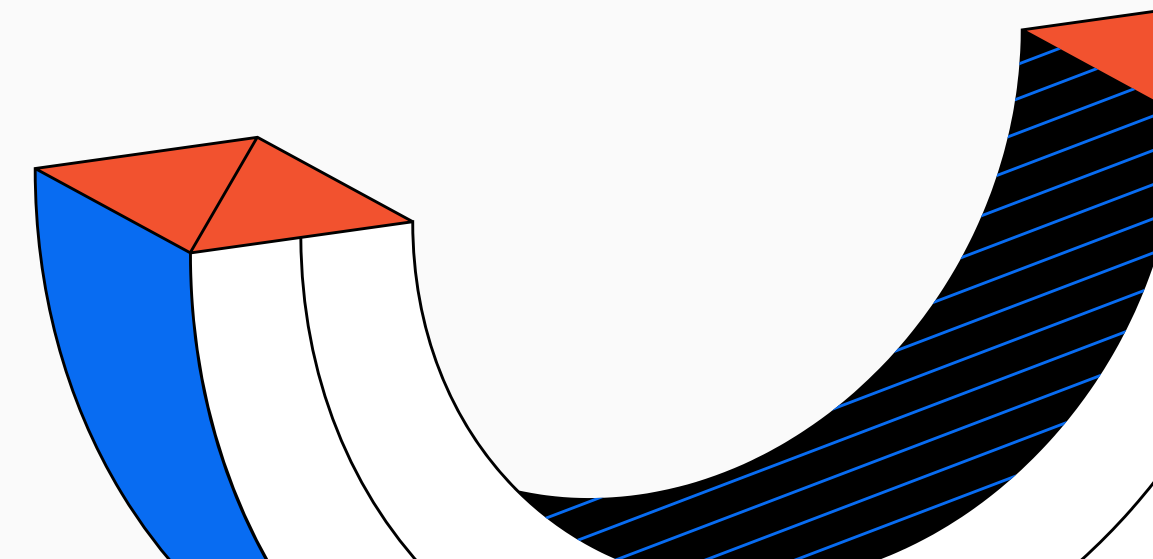
Background

More than 100 Millions of Job Application was sent to Jobstreet in 2021. (JobStreet Report, 2021)

And HR said candidates only stand chance 30,89% of getting hired, than receiving an offer. And 1 of 6 of them are rejecting offer. (Glassdoor, 2021.



[Back to Agenda Page](#)





NYC job-seekers will soon be guaranteed salary estimates

[Economy](#) Oct 31, 2022 1:31 PM EST

NEW YORK (AP) — Starting this week, job-seekers in New York City will have access to a key piece of information: how much money they can expect to earn for an advertised opening.

HR Headaches: Candidates Aren't Accepting My Job Offers

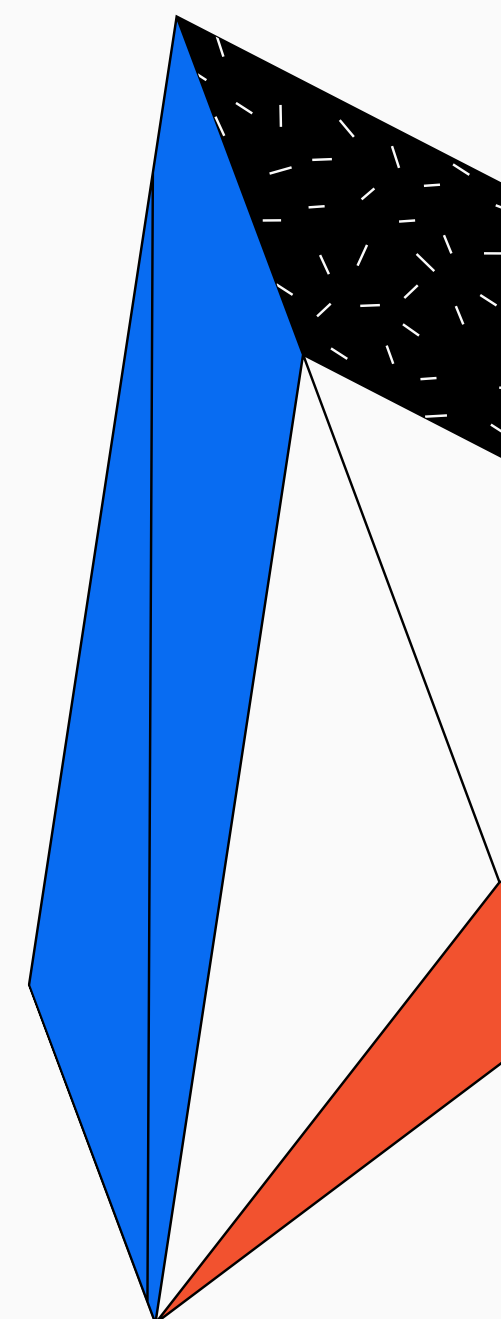
Whether candidates are ghosting or rejecting your job offers, you're not alone. In 2020, Glassdoor reported that 1 in 6 offers was rejected on average. In some industries, that rate was considerably higher.

U.S. workers have wasted millions of hours applying to jobs with the wrong salary—how to avoid it

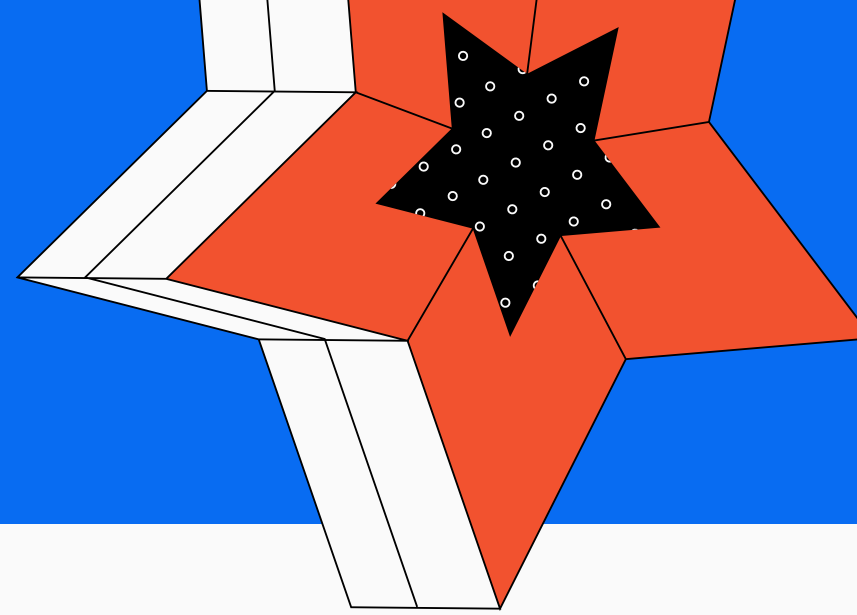
 Ashton Jackson
[@ASHTONLINNELL](#)

SHARE    

Gaji menjadi salah satu faktor **penentu** apakah lowongan itu **berhasil didapatkan oleh pelamar/** apakah **employer** mendapatkan **kandidat**



Reason for Choosing



[Back to Agenda Page](#)

19%

**REJECT THE
OFFERING**

Because salary isn't
right for them

73%

**HR Hope Salary
Negotitations**

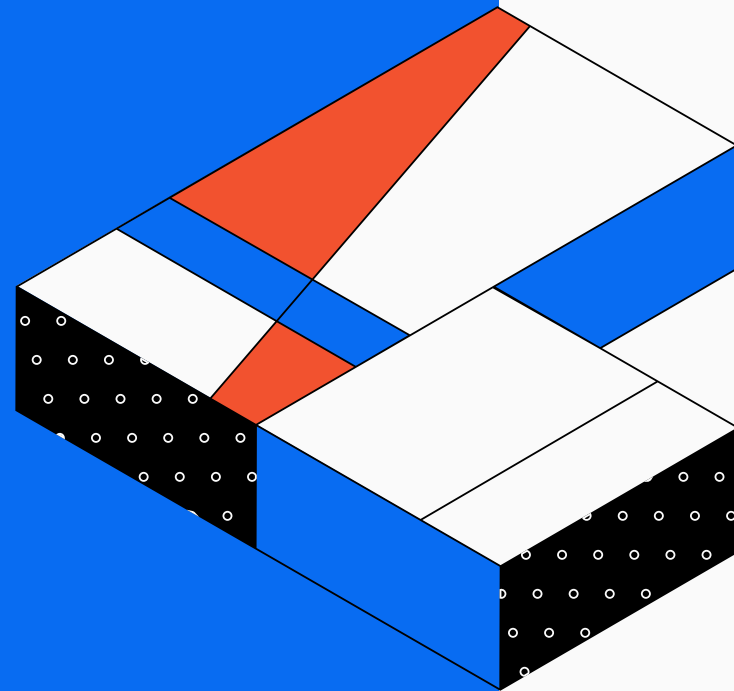
To improve
performance in the job

55%

**Candidates
Won't Negoitate**

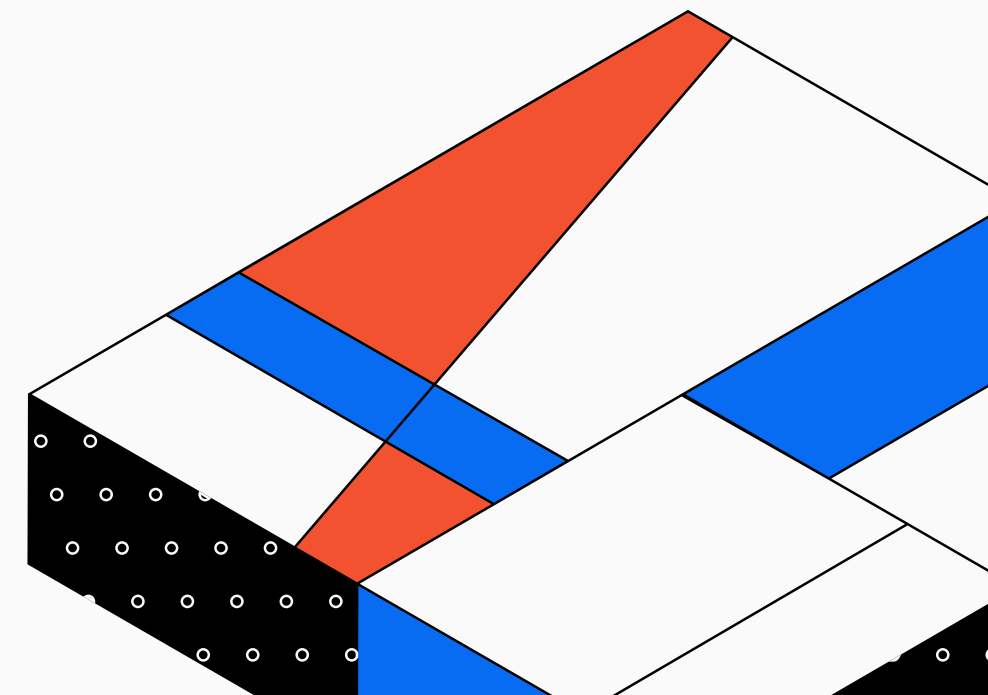
Because **they don't know
how much salary is right
for them**

Job Description and Salary in Indonesia



[Back to Agenda Page](#)

- Job_Function
- Company_Size
- Location
- Career_level
- Job_benefit
- Job_description
- Experience_level
- Company_industry
- Employment_type
- **Salary**



Data Preprocessing

[Back to Agenda Page](#)

Cleansing

- Handling Missing Values
Categoricals features : Modus
Numeric feature : Media
- Duplicated Data is left as it is because there is no column that described it is from a single source

Regrouping

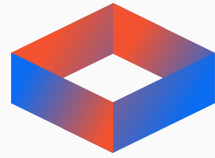
- We are regrouping column with unique values to some class in job_function and career_level

Drop Value with Lots of Unique

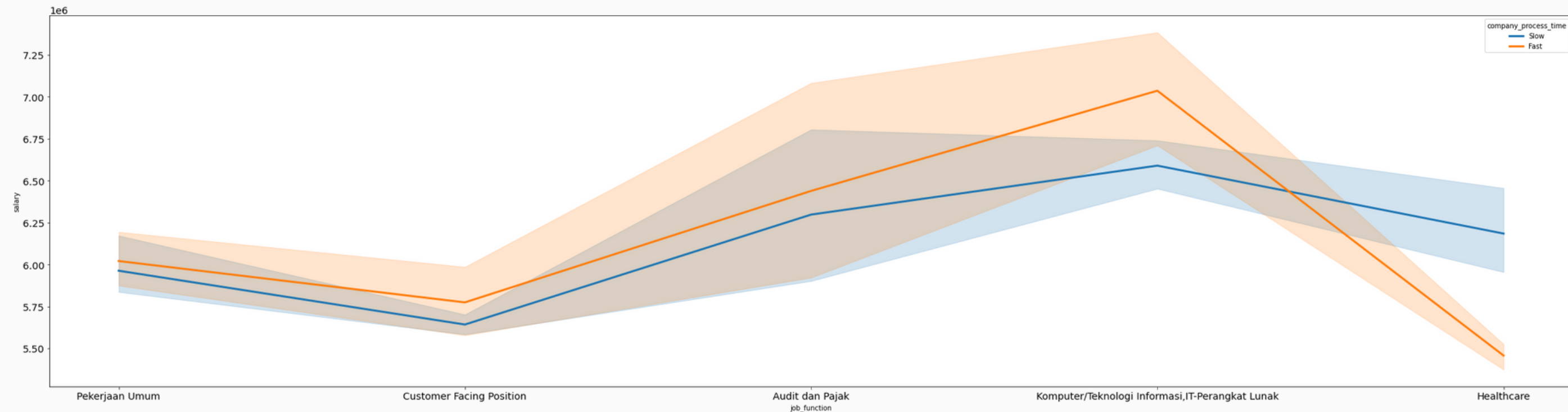
- We are dropping job_title, job_benefit and job_description because there is a lot of unique data

Feature Engineering

- Label Encoding for location, education, company size, process time, company industry, experience and career.
- One Hot Encoding for employment type and job function.
- Feature Scalling = Min Max Scaller for Salary

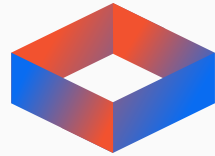


EDA

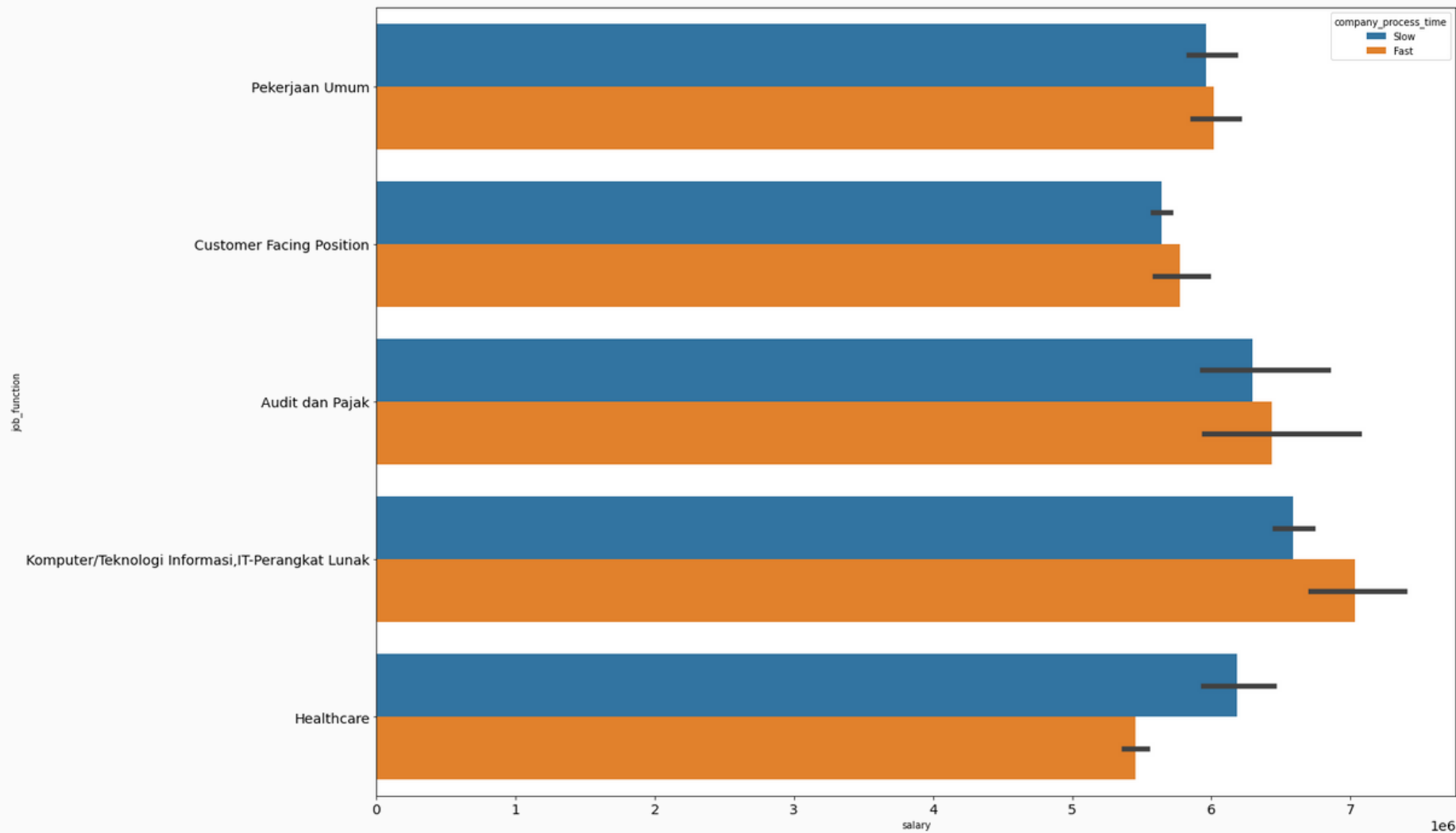


From this Data, we can conclude that no class in the column **location, job_function, and career_level** has very little difference with the mean of the salary. Therefore, we have to rely on another variable to decide salary difference

We suggest to **add more data regarding job with more salary difference** to this dataset to give the ML more information to tell exact salary for each positions. But we can conclude, **IT Positions tend to get higher salary** than non-IT Positions



EDA

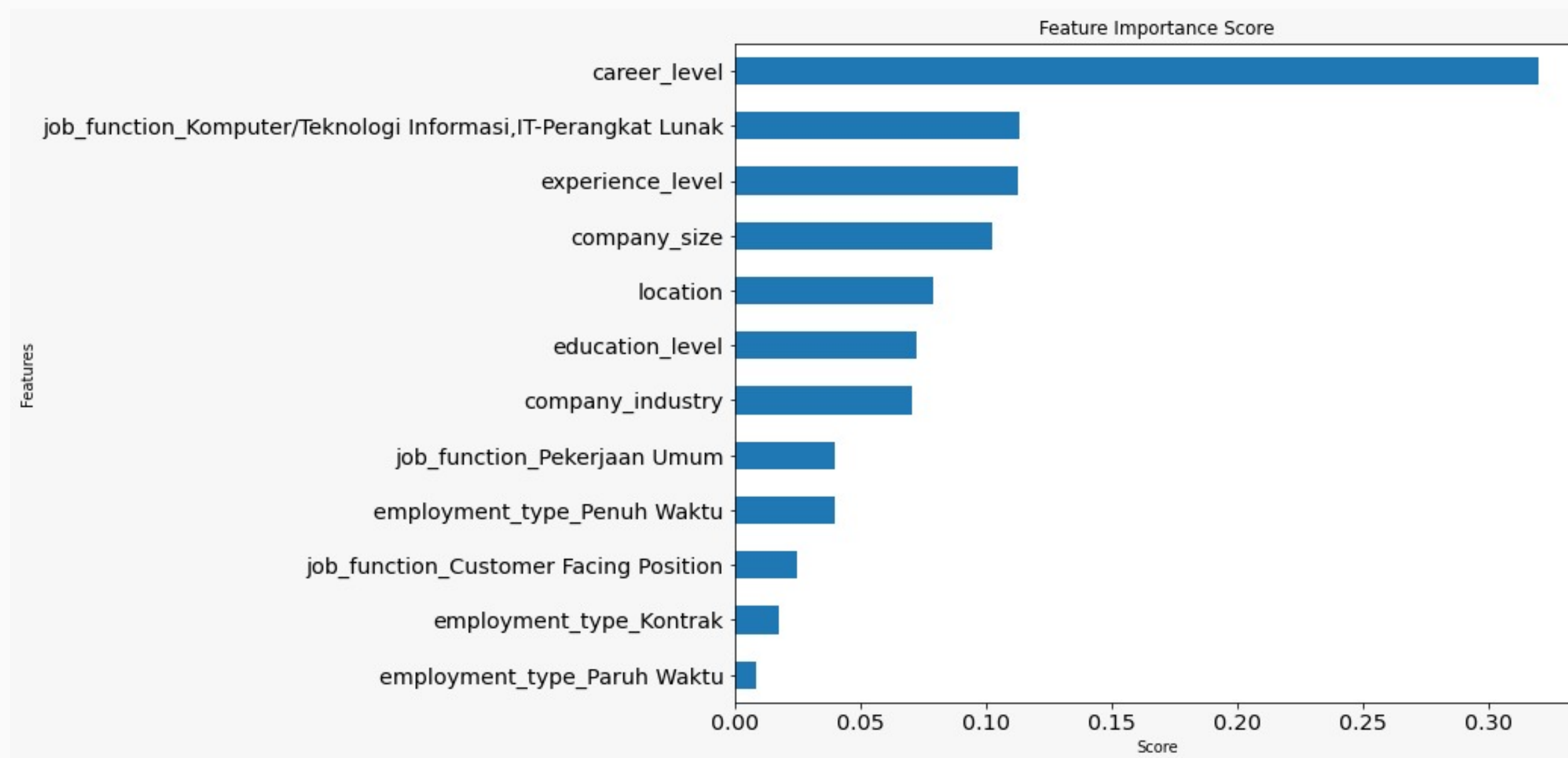


From a **recruitment process** point of view,
There is **a tendency to have more salary if the company is recruiting faster**. It is probably due to urgent vacant positions that needed to be filled, **except for the healthcare sector**.

[Back to Agenda Page](#)



Feature Importance



As mentioned before, **IT Positions tend to get more salaries and it is detected as an important feature** for this prediction model. Other than that, **career level, experience, company size, education , industry also some job_function is also important** part to predict the salaries.

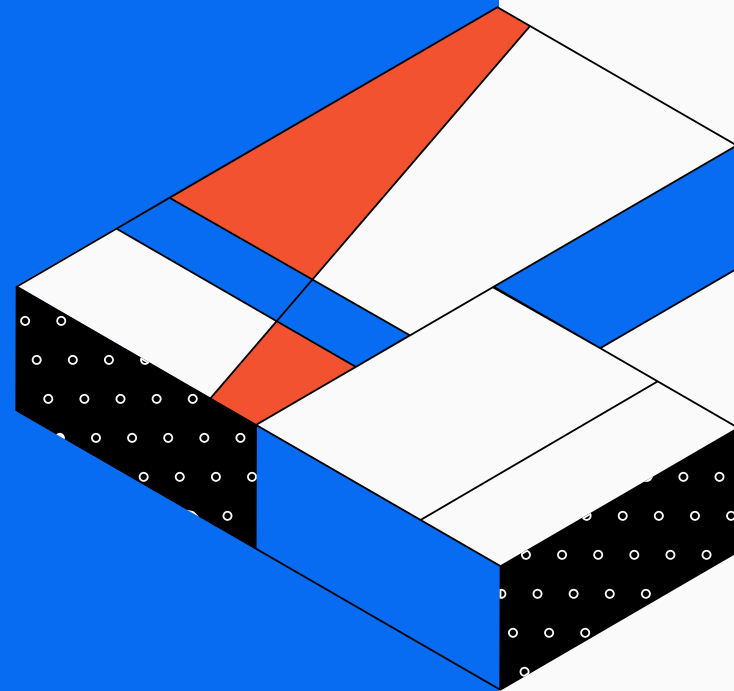
Feature Engineering

1. Label encoding: company process time, experience level, education level, career level, company size, and salary currency
2. One Hot Encoding: location, employment type, job function, and company industry

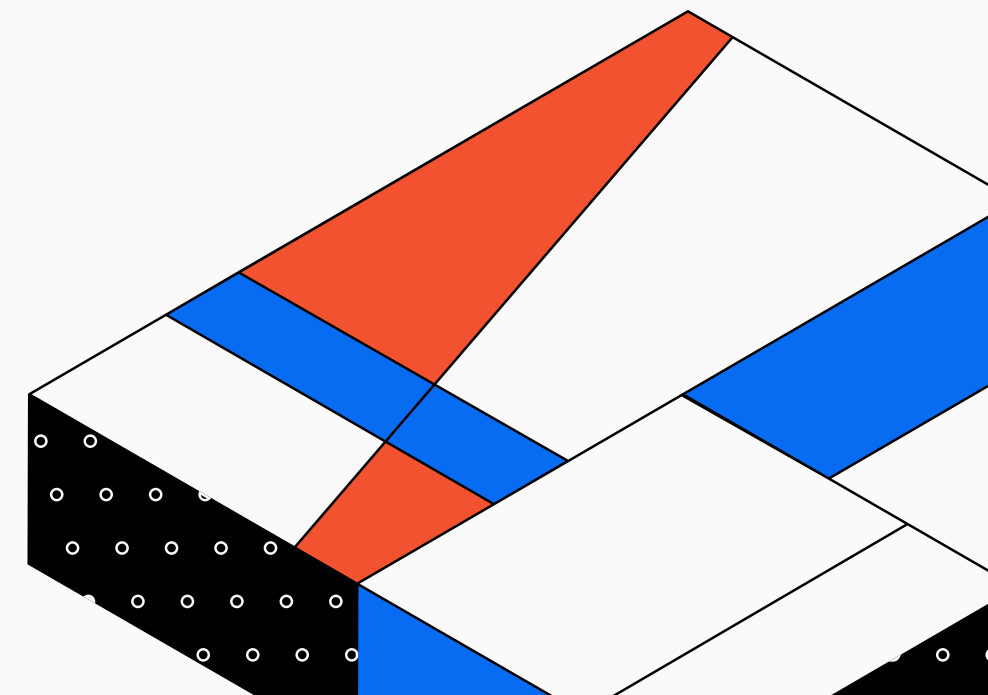


Modeling

- Split dataset : 70% train and 30% test
- Model Machine Learning : Linear Regression, Decision Tree, Random Forest, and XGBoost Regressor



[Back to Agenda Page](#)

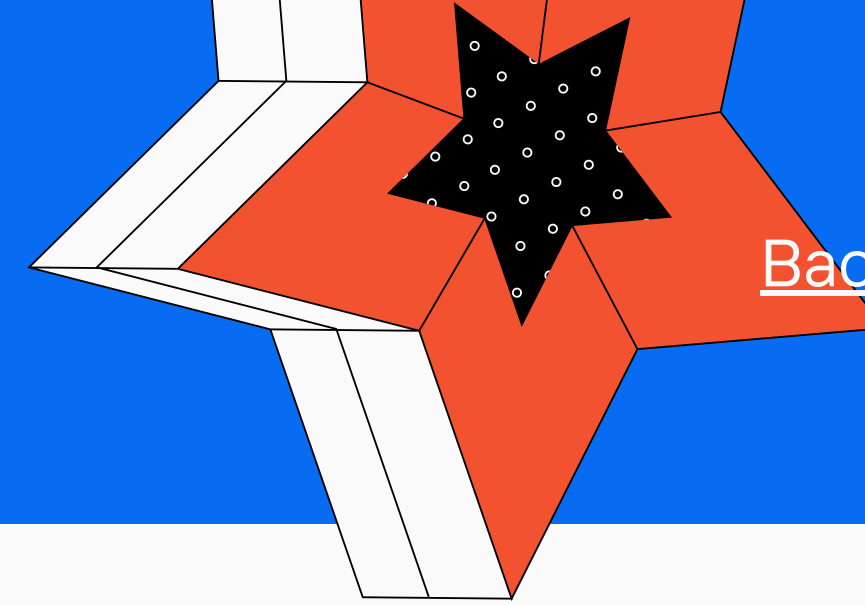


Model Comparison

Algorithm	RMSE	MAE	R2	Train	Test
Linear Regression	0.061	0.0290	0.1001	0.091	0.1005
Decision Tree	0.0587	0.0270	0.1601	0.1601	0.0954
Random Forest	0.0588	0.0272	0.1584	0.1584	0.1019
XGBoost	0.0589	0.0274	0.1540	0.1540	0.1059



Business Recommendation



[Back to Agenda Page](#)

1. **The datasets need new data with variation range of salaries**, so there will add new learning patterns for the ML to know which factor defines more in salary offering. Is it location, position, urgency or else?
2. Or, **we can deploy the dataset per industry** that we are offering to predict salaries, so the function won't get mixed by each other and **the result will be more specified** depend on the industry, then the positions.
3. To implement this, **usually job_title and job_descriptions plays an important role**, but due to more than 20k unique value, **we need a new method** so job_title can be used but won't cause mispredictions. Using NLP might be a quite catch!

Launch Plan

Divide implementations by Industry(MVP)

1. We will get this implemented on web, but first we need to simplify the logic of the predictors in order to get more accuracy and context of salaries.
2. Usually same Role in Different industry pays different too, so as MVP we just have to predict few title in 1 industry

Gather more data that represent the Position More

- In tech industry, we can't just predict the salary based on the "IT Position".
- There is also more position like UI Design, Product, Data with each of its Career level with the name "Associate", "Principal" and more and that is more defining the salary than else

Train NLP

- After we could get more data to train, we can get job titles or even job description to play a role in defining salaries.
- This way, CV Scanner won't be a problem to deploy its full potential

Full Launch

- After we trained the model we can use Flask as CV Reader and script producer and then connect it to trained ML with new datasets

Reference



<https://www.kaggle.com/datasets/canggih/jog-description-and-salary-in-indonesia>

<https://www.cnnindonesia.com/ekonomi/20201118154433-532-571422/pencari-kerja-di-indonesia-tembus-69-juta-orang-per-tahun>

<https://www.cnbc.com/2020/01/28/half-of-job-seekers-rejected-a-job-offer-after-an-interviewwheres-why.html>

<https://www.careerplug.com/blog/reasons-candidates-turn-down-job-offers/>

www.recruitingnewsnetwork.com/posts/job-seeker-nation-fear-uncertainty-and-a-serious-desire-to-work-remotely-dominate