



**IRON  
HACK**



# SkillScanner



**Authors:** Yu Ting Hu Wu  
Mónica Graciela Duarte Arrieche  
**Date:** 29-Sep-2022



# Table of Contents

1

Why SkillScanner?

2

Technology Stack

3

Project Overview

4

Exploratory Data Analysis

5

Predictive Modeling

6

SkillScanner User Interface





```
try:
    print('Write your LinkedIn URL:')
    URL = input()
    if URL[-1] == '/':
        url_name=URL.split('in/')[1].split('/')[0]
    else:
        url_name=URL.split('in/')[1]

# Calling the LinkedIn API
profile = api.get_profile(url_name)
skills_profile = api.get_profile_skills(url_id=profile['entityid'],split('urn:li:profile:')[1])
skill_list = [dicc['name'] for dicc in skills_profile]
dicc_skills = {key: 0 for key in list(job_data.columns)}

# Generating the wordcloud
print('these are your skills:')
text = ' '.join(map(str, skill_list))

d = path.dirname(__file__) if '__file__' in locals() else os.getcwd()
cloud_mask = np.array(Image.open(path.join(d, 'oval.png')))
mask = np.array(Image.open('oval.png'))
wordcloud = WordCloud(background_color='white', max_words=1000, mask=mask).generate(text)

plt.savefig('N.png')

plt.figure(figsize=(15,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.savefig('wordcloud.png')
plt.show()

# Preprocessing the data
print(' ')
print('-----')
print('Your skills among the most popular in the data industry:')
for i in list(job_data.columns):
    if i.lower() in text.lower():
        dicc_skill[i] = 1

# Individually find R and C skills as one word
dicc_skill['R'] = np.where(np.find_only_word('r', text), list(1), list(0))
dicc_skill['R'] = int(dicc_skill['R'])
dicc_skill['C'] = np.where(np.find_only_word('c', text), list(1), list(0))
dicc_skill['C'] = int(dicc_skill['C'])
dicc_skill['ML'] = np.where(np.find_only_word('ml', text), list(1), list(0))
dicc_skill['ML'] = int(dicc_skill['ML'])

for key, val in dicc_skill.items():
    if val==1:
        print('+' + key)

if sum(list(dicc_skill.values())) < 4 :
    print('Sorry!')
    print('It looks like you don't have enough data skills yet. Check your profile')
    return

else:
    job_data = job_data.append(dicc_skill, ignore_index=True)
    production = job_data.tail(1)

# Using the model to predict
target = 'Job Category'
X_prod = production.drop(target, axis=1)
y = production[target]

y_pred = final_model.predict(X_prod)
final_model_predict_proba(X_prod)

results = pd.DataFrame(final_model_predict_proba(X_prod), columns = ['Data Analyst', 'Data Engineer', 'Data Scientist'],)
```

POTLIGHT ON BIG DATA  
**spotlight**

ARTWORK Tamar Cohen, Andrew J Buboltz  
2011, silk screen on a page from a high school  
yearbook, 8.5" x 12"

## Data Scientist: The Sexiest Job of the 21st Century

Meet the people who  
can coax treasure out of  
messy, unstructured data.  
by Thomas H. Davenport  
and D.J. Patil



70 Harvard Business Review October 2012



## ● Programming Language

- ✓ Python

## ● Libraries

- ✓ Numpy
- ✓ Pandas
- ✓ Pickle
- ✓ Regex

## ● Data collection

- ✓ API REST - LinkedIn
- ✓ Web Scrapping – Glassdoor

## ● EDA & Visualization

- ✓ Pandas Profiling
- ✓ Plotly Express

## ● Predictive Modeling

- ✓ Scikit-Learn



## SkillScanner Project RoadMap

### Identify skills

Identify role skills, which will serve as input for supervised classification algorithms.

1

### EDA

Get data from job offers data and make data-driven insights by means of EDA

2

### ML Modeling

Build Machine Learning models to make a prediction of how a profile fits in each of the job roles

3

### SkillScanner

Develop a product that is able to:

- ✓ Scan skills from a LinkedIn profile.
- ✓ Identify in-demand skills in Data.
- ✓ Predict its best job role fitting.

4



Fig.1 Total job offers per category

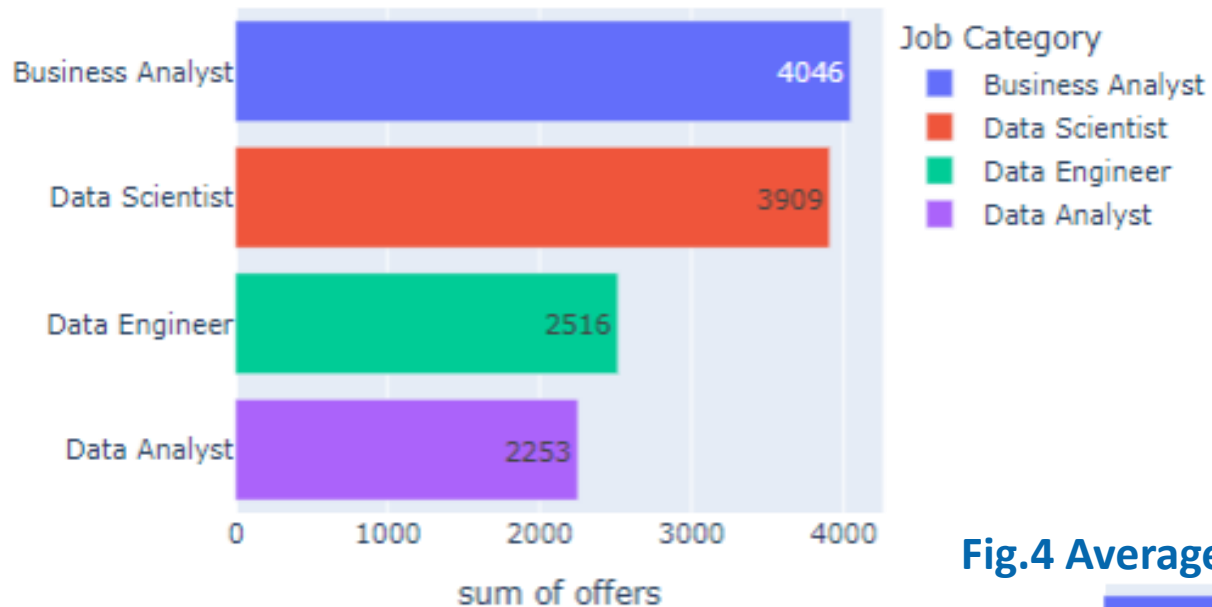


Fig.2 Total job offers per category (%)

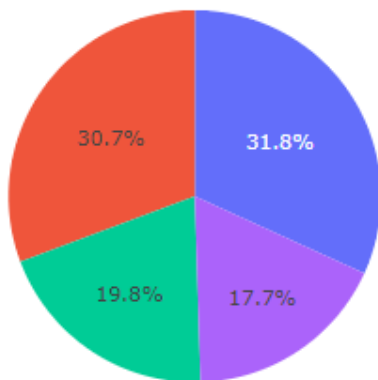
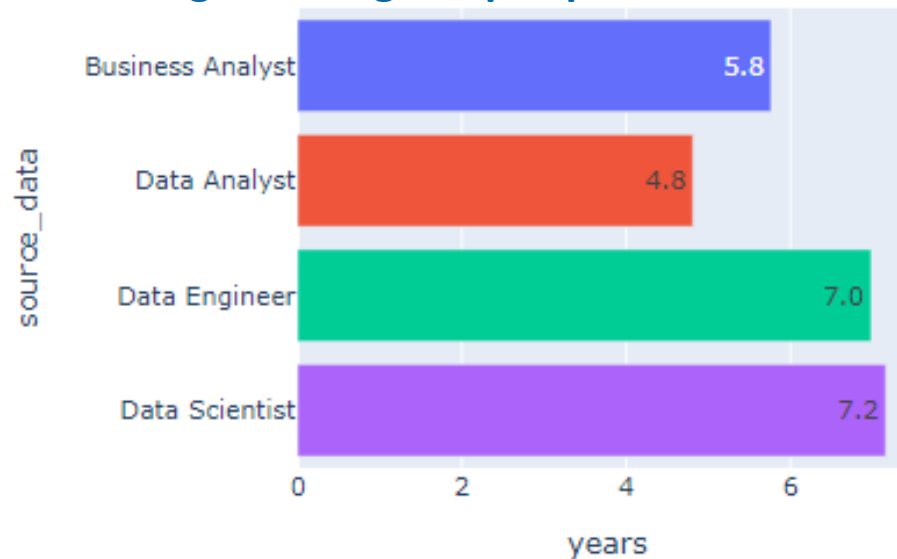


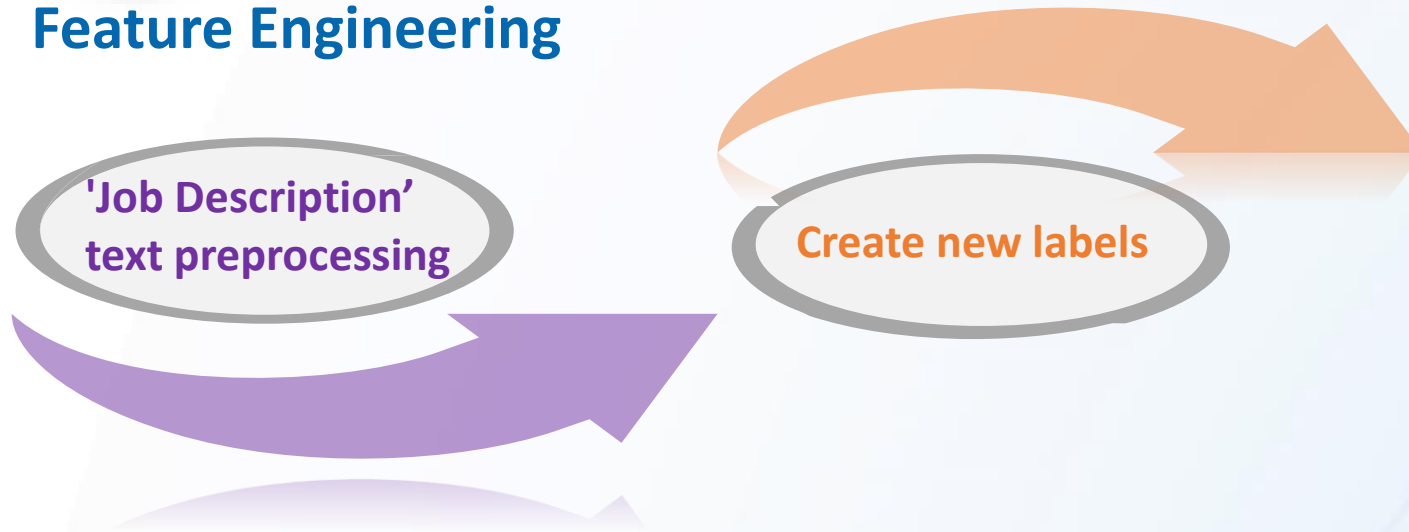
Fig.3 Salary range (min-max) per job category



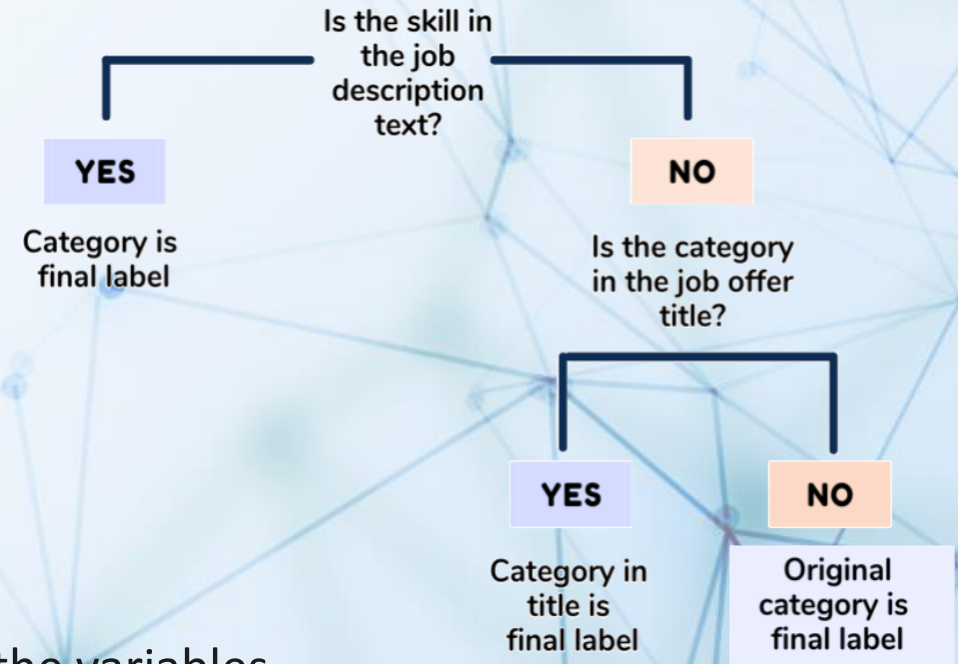
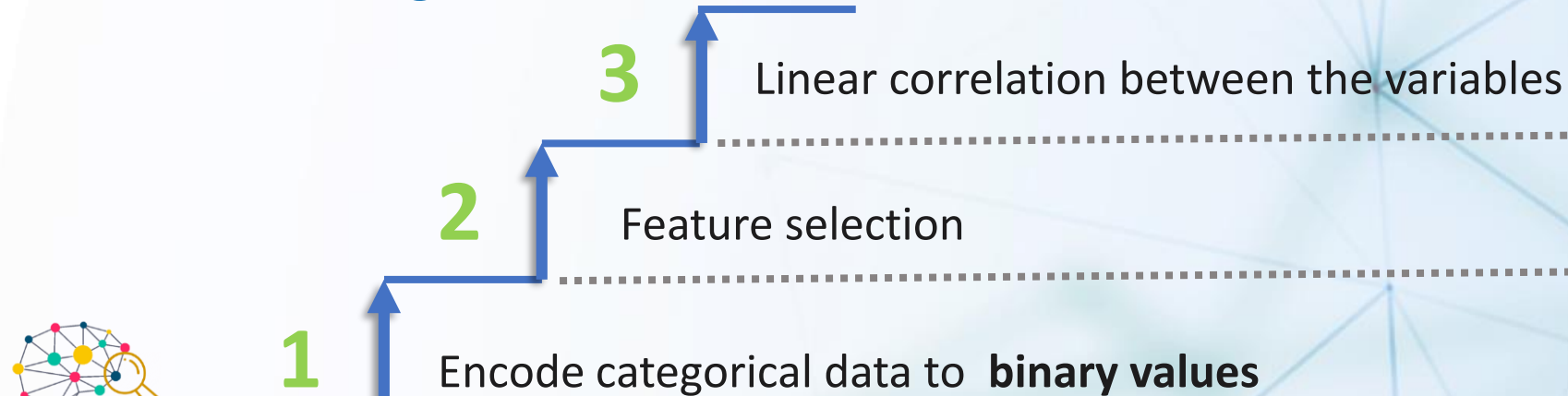
Fig.4 Average req. experience



## Feature Engineering



## Feature Encoding



## Machine Learning Algorithms





## Classification models process



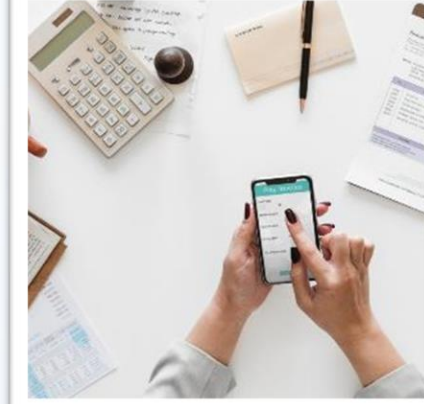
### Feature Engineering

- ✓ Label Assignment
- ✓ Feature Encoding
- ✓ Feature Importance
- ✓ Feature Selection



### Preprocessing

- ✓ Split data (80-20)
- ✓ Check data balance
- ✓ Feature Scaling



### Hyperparameter tuning

- ✓ Grid Search
- ✓ Cross Validation
- ✓ Multiclass parameters



### Training & Testing

- ✓ Model Training (80)
- ✓ Model Testing (20)
- ✓ Production simulation with unknown data



### Evaluation

- ✓ Evaluation metrics selection
- ✓ Results comparison and evaluation
- ✓ Confusion Matrix

# Predictive Modeling

Accuracy	Precision	Recall	f1	Set	Model
0.644505	0.644505	0.644505	<b>0.644505</b>	test	<b>Logistic Regression</b>
0.650856	0.650856	0.650856	<b>0.650856</b>	train	<b>Logistic Regression</b>
0.634337	0.634337	0.634337	0.634337	test	Knn
0.677359	0.677359	0.677359	0.677359	train	Knn
0.647243	0.647243	0.647243	0.647243	test	PCA + Logística
0.650367	0.650367	0.650367	0.650367	train	PCA + Logística
0.678138	0.678138	0.678138	0.678138	test	Random Forest
0.718924	0.718924	0.718924	0.718924	train	Random Forest
0.688307	0.688307	0.688307	<b>0.688307</b>	test	<b>Gradient Boost</b>
0.753350	0.753350	0.753350	<b>0.753350</b>	train	<b>Gradient Boost</b>
0.678530	0.678530	0.678530	0.678530	test	XGB
0.722054	0.722054	0.722054	0.722054	train	XGB

With SkillScanner, you can:

- ✓ Scan all skills from any LinkedIn profile
- ✓ Identify in-demand skills
- ✓ Predict what job role fits you best!

... in less than **30 seconds!**

