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

Why SkillScanner?

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

Technology Stack

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

Project Overview

SkillScanner Project RoadMap

Identify skills

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

EDA

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

ML Modeling

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

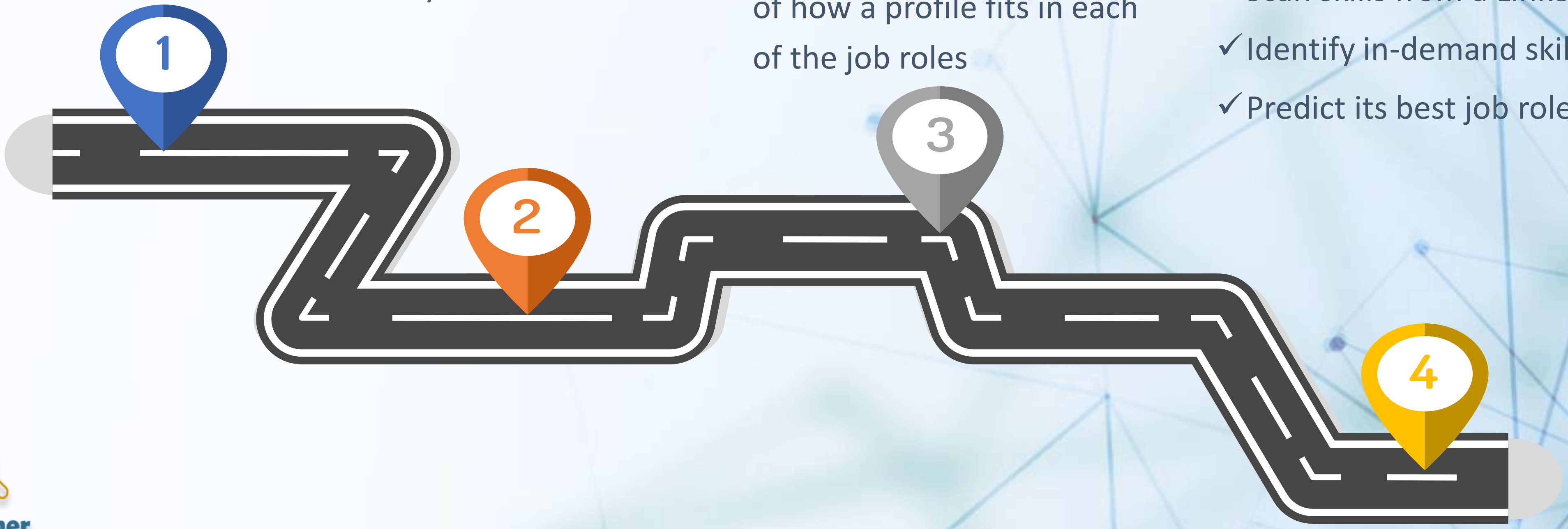
SkillScanner

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.



SkillScanner

Exploratory Data Analysis

Fig.1 Total job offers per category

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

Fig.2 Total job offers per category (%)

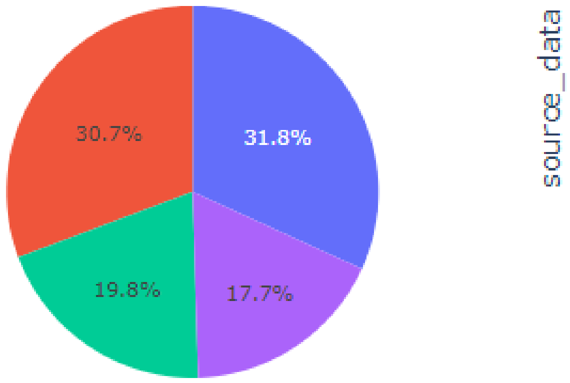
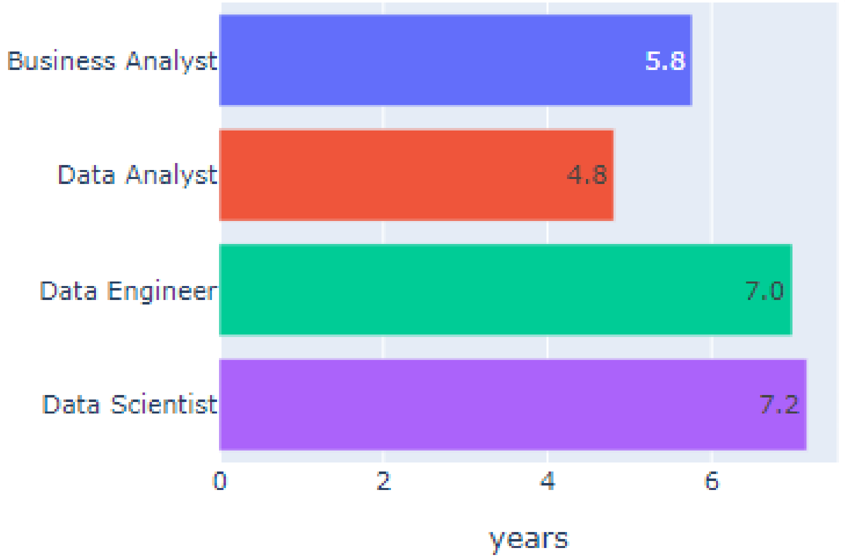


Fig.4 Average req. experience



SkillScanner

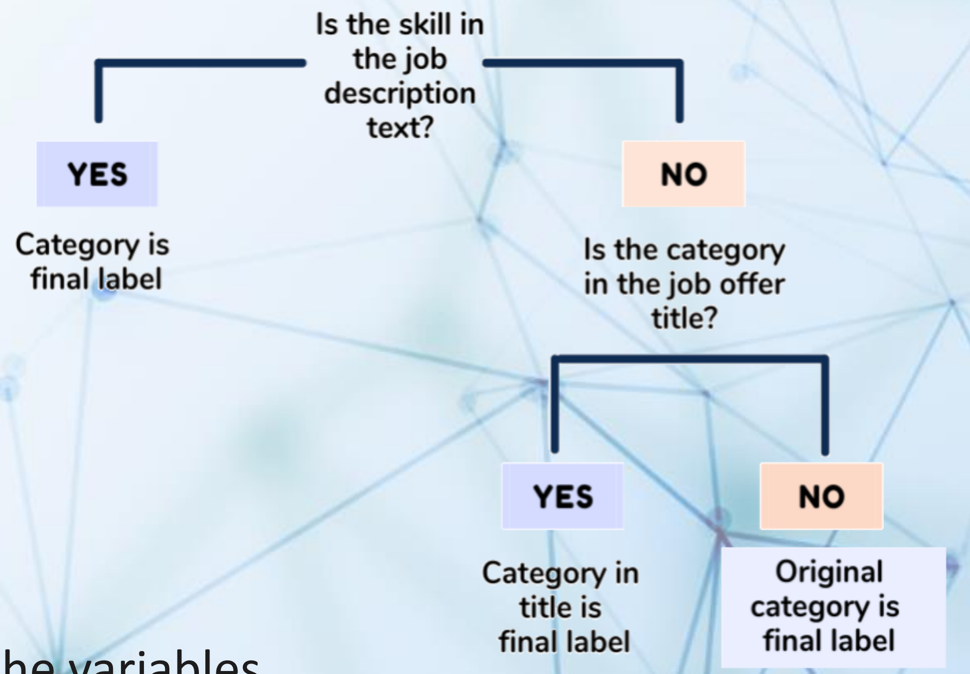
glassdoor

Predictive Modeling

Feature Engineering

'Job Description' text preprocessing

Create new labels



Feature Encoding

3 Linear correlation between the variables

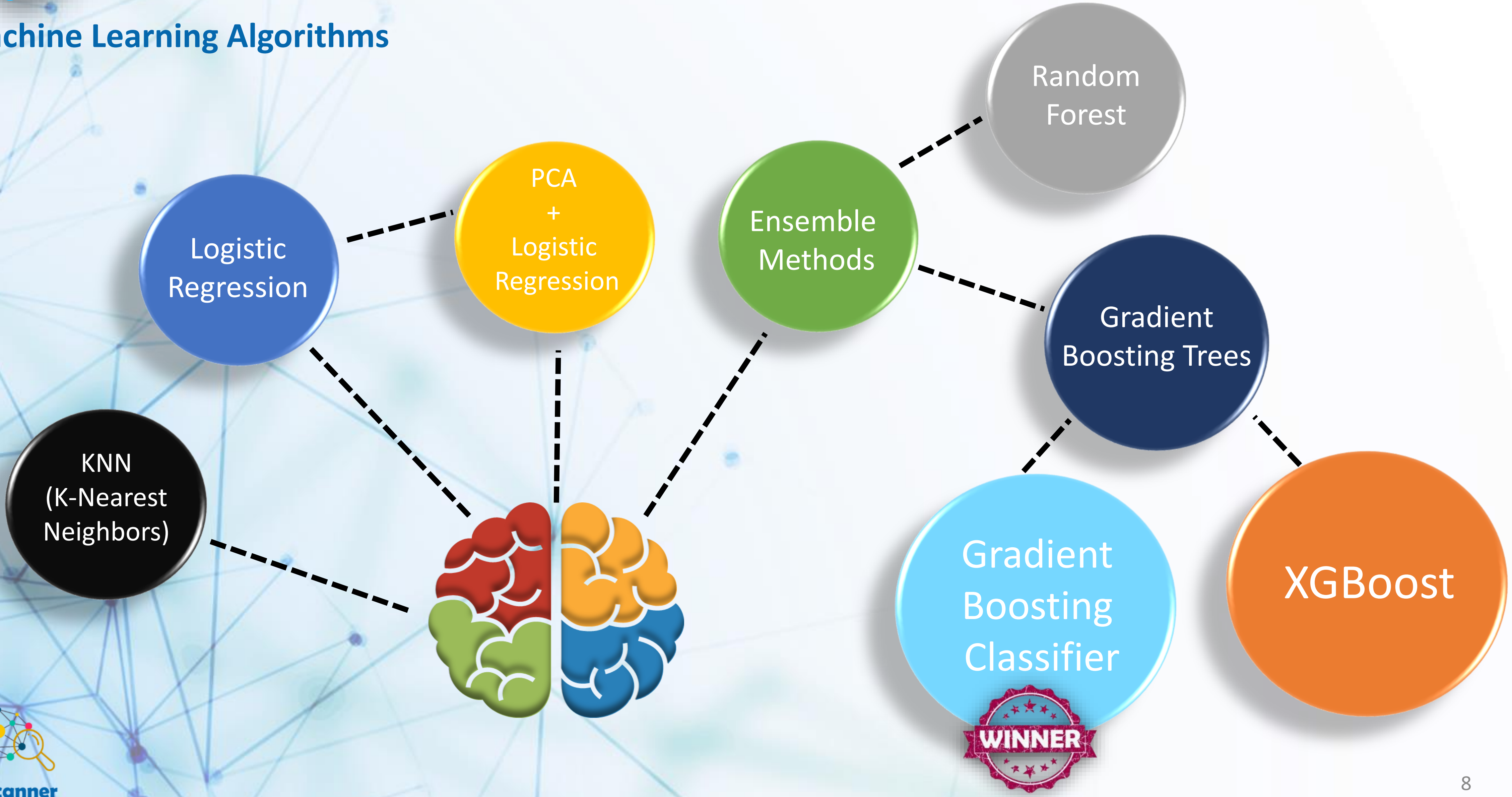
2 Feature selection

1 Encode categorical data to **binary values**

SkillScanner

Predictive Modeling

Machine Learning Algorithms



Predictive Modeling

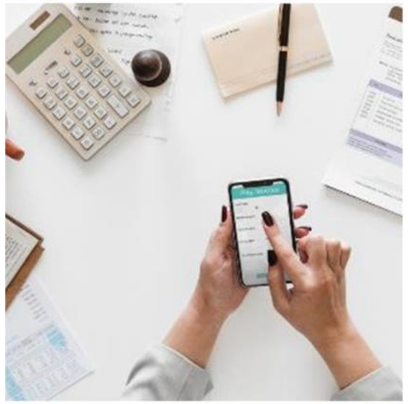
Classification models process

Feature Engineering

- ✓ Label Assignation
- ✓ 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

SkillScanner

Predictive Modeling

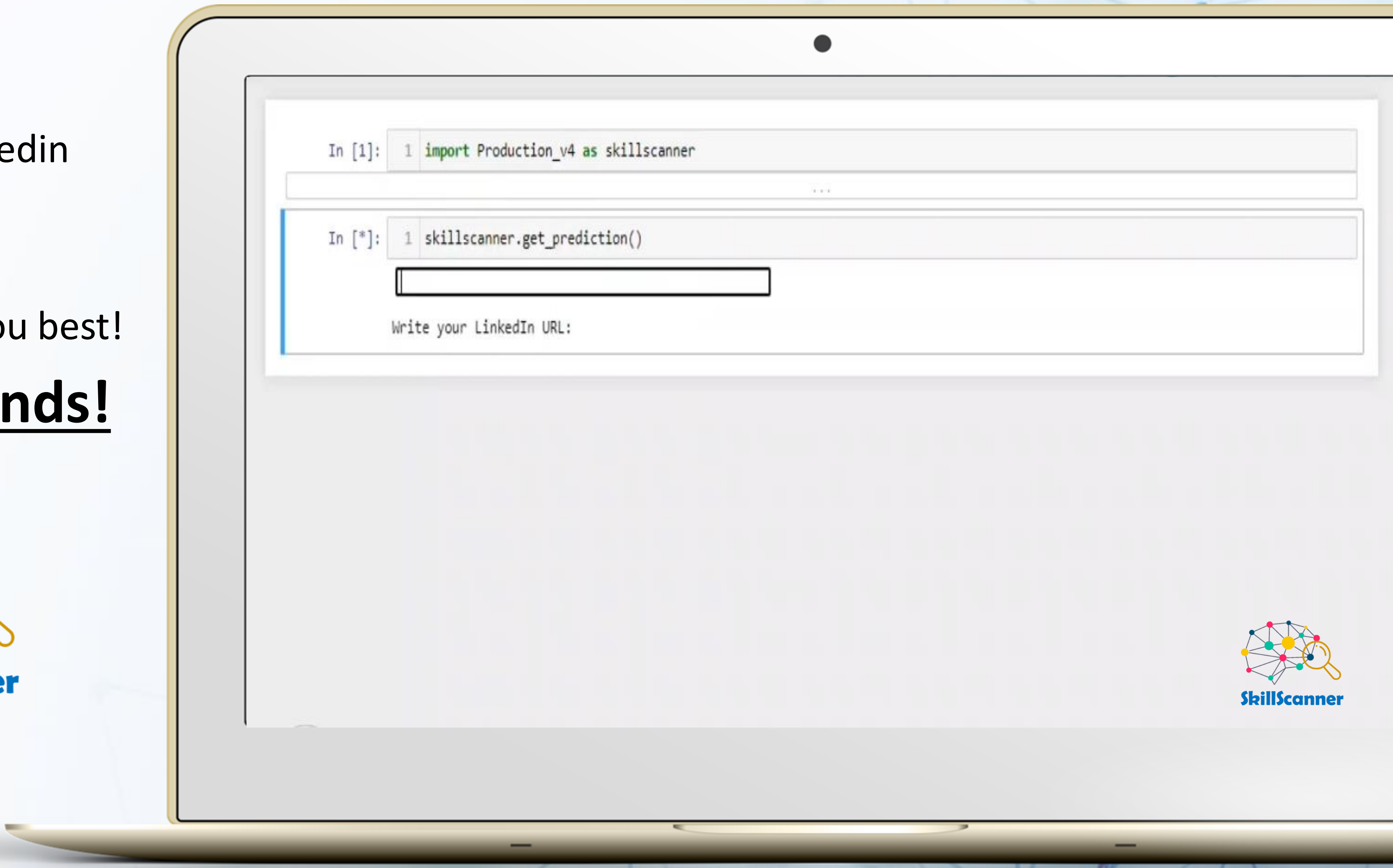
Accuracy	Precision	Recall	f1	Set	Model
0.644505	0.644505	0.644505	**0.644505**	test	**Logistic Regression**
0.650856	0.650856	0.650856	**0.650856**	train	**Logistic Regression**
0.634337	0.634337	0.634337	0.634337	test	Knn
0.677359	0.677359	0.677359	0.677359	train	Knn
0.647243	0.647243	0.647243	0.647243	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	**0.688307**	test	**Gradient Boost**
0.753350	0.753350	0.753350	**0.753350**	train	**Gradient Boost**
0.678530	0.678530	0.678530	0.678530	test	XGB
0.722054	0.722054	0.722054	0.722054	train	XGB

SkillScanner User Interface

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!**



Q & A

QUESTIONS?
THANK YOU
WHEN?
WHERE?
WHAT?
HOW?
WHO?
Where?
What?
When?
Why?