Business Analytics 2022

PREDICTING JOB ACCEPTANCE DECISION FOR BETTER CANDIDATE SELECTION

WORK BY:

Diogo Valente
up201806473

Filipe Azeredo
up201806315

Inês Santos
up201806346

Jesse Purkamo
up202111212

Joana Pina
up201806335

João Silva
up201806335

Margarida Sá
up201806662

Otto Veijalainen
up202111487

What's Job Change about?

Business Context

A company wants to hire data scientists among people who successfully pass some courses conducted by the company. They want to know which of these candidates really want to work for the company after training and who will look for another employment

Goals

Predictive Model for the Prob. of Candidate Looking for New Job

Identify **Key Features** Affecting Employee Decision

The Job Change Dataset

19 158
Data Entries
(per employee)

12
Features
(+ ID and Target)

9
Categorical Feat.
(w/ high cardinality)

+9%
Missing Values
(20 733)

44
Uncoherent Obs.
(University & Primary School)

25%
Positive Target
(High imbalance)

Main Datasets were created with general Pre-Processing

Missing & Noisy Data

- ✓ ['Major_Discipline'] New category "No Major" for Primary & High School with no university enrollement (**1195 obs**)
- ✓ Removed observations with 4+ N/A's (**1.30%**)
- ✓ Removed observations: Enrolled University with Primary School (**44 obs**)
- ✓ Fill in with **kNN** and **MICE** techniques

Dealing with Categorical Variables

- ✓ City Feature: **Target Encoding**
- ✓ Ordinal Features: **Label Encoding** with numerical labels
- ✓ Nominal Features: **One-Hot Encoding** (dummy variables with linear dependence), while dropping redundant variables
- ✓ **Clustering One-Hot Encoding**: the dummy variables were clustered to reduce dimensionality

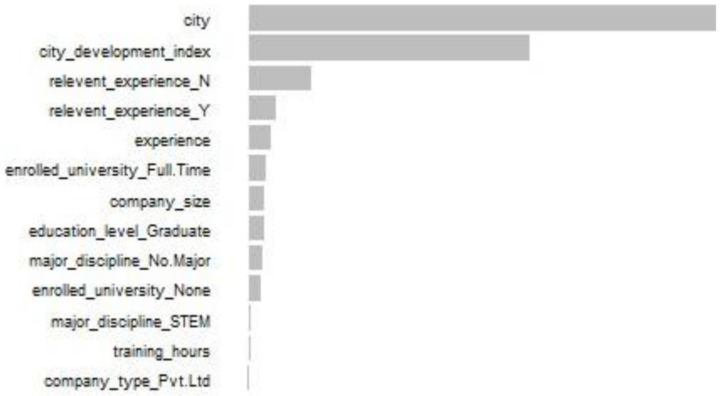
✓ Methods that generate multiple datasets

Different Feature Selection techniques were used and tested for better results

Correlation Analysis

- ✗ **Point-Biserial Correlation**: between Numeric Features and Target
- ✗ **Chi-Square Test**: between Categorical Features and Target
- ✓ **Polychoric Correlation**: between Ordinal Features
- ✓ **Cramer-V Test**: between Nominal Features
- ✓ **Boruta**: for Random Forest Model

Feature Importance



9 Classification Techniques Were Used

1R

OneR

NB

Naïve Bayes

Logistic Regression

Decision Tree

Random Forest

kNN

XGBoost

LightGBM

Neural Network

Our modelling process had 5 main steps

1 Dataset Selection & Preparation

- ✓ Normalization
- ✓ Feature Selection

2 Data Balancing

- ✓ Oversampling
- ✓ Undersampling
- ✓ Both & Normal
- ✓ ROSE
- ✓ SMOTE

3 Parameter Tunning

4 10-Fold Cross Validation

5 Prediction and Metrics Evaluation

We stuck with 3 Key Evaluation Metrics Better Suited For Class Labels

01 F1 Score

Combines precision and recall

Ideal for Imbalanced Classification

02 Accuracy

Number of correctly predicted data points out of all data points

Not Ideal for Imbalanced Classification

03 G-Mean

Balances the classification performances on both the majority and minority classes

Ideal for Imbalanced Classification

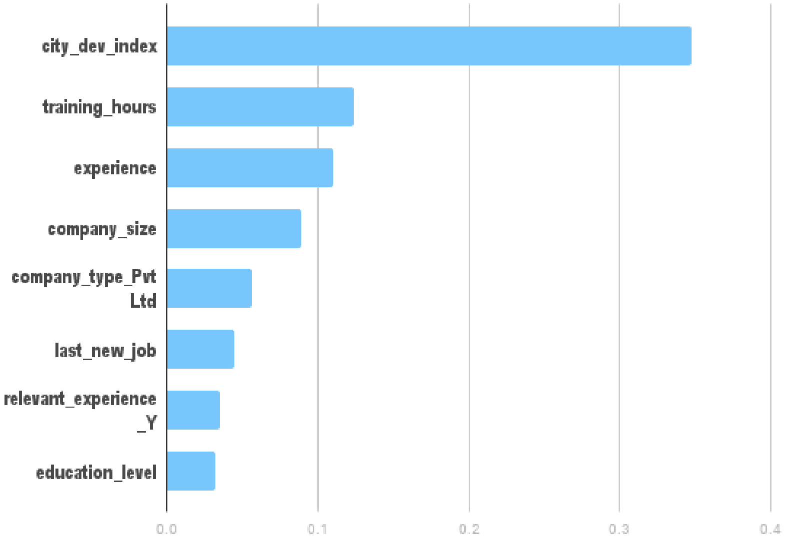
Source: [Machine Learning Mastery]

From all models, we dug deeper into the best 4

Model	F1	Acc.	GMean	Proc. Time	Sampling	Dataset Description
Decision Tree	86.44	79.10	66.59	11 min	Normal	MICE filling, One Hot, Removed Highly Target Correlated Vars
Random Forest	**87.42**	**80.60**	69.00	2 min	Normal	MICE filling, One Hot
XGBoost	87.10	79.96	**72.49**	**< 2min**	Normal	MICE filling, One Hot, Normalized, 'City' Removed
Neural Net	83.62	76.82	68.19	> 30 min	Over	MICE filling, One Hot, Normalized

XGBOOST model seems to be the most promising

XGBoost Feature Importance



Tunned Hyperparameters

7 ⇄ **max_depth**
Max depth per tree

0.1 ⇄ **eta**
Learning rate

100 ⇄ **nrounds**
No. of trees/ boosting rounds

F1 87.11%

Accuracy 79.99%

G-Mean 72.37%

With the XGBoost model information, the company can improve its recruitment strategy

Better City Targeting
With city-specific course marketing and advertising

Training Hours Thresholds
Only accept candidates with a certain no. of training hours

Job Survey in Courses
Survey for experience, years since last job and size/ type of current

Future Model/ Data Improvements

Candidate to Employee Tracking
Build course candidate database for tracking after employment (for better data)

- **Time Until Leaving Company**
- **Types of Courses Taken**
- **Course Grades**

Thank you for your attention!

Room for your feedback