Predicting Hiring Decisions in Recruitment Data

Christoforos Zisis

Machine Learning
Professor: Michail G.Lagoudakis
MSc. Machine Learning and Data Science

Technical University of Crete, Spring 2024

Motivation

- Data from recruitment team provides information to train the Machine Learning models.
- Collected the candidate's features.
- The model predict if the candidates is qualified for hiring.

Methodology

1. Automating Machine Learning Pipelines
2. Improvement of Automating Machine Learning Pipelines
3. Candidates' Features And Prediction
4. Conclusions
5. References

Phase 1
Automating Machine Learning Pipelines

Description of Dataset

Predicting Hiring Decisions in Recruitment Data:

This dataset provides insights into factors influencing hiring decisions. Each record represents a candidate with various attributes considered during the hiring process.The goal is to predict whether a candidate will be hired based on these attributes.

The Data includes:

- Records: 1500
- Features: 10
- Target Variable: Hiring Decision (Binary)

Description of Parameters / Methods

- Hyper-parameter Tuning
- Ensemble learning
- Automating Machine Learning Pipelines

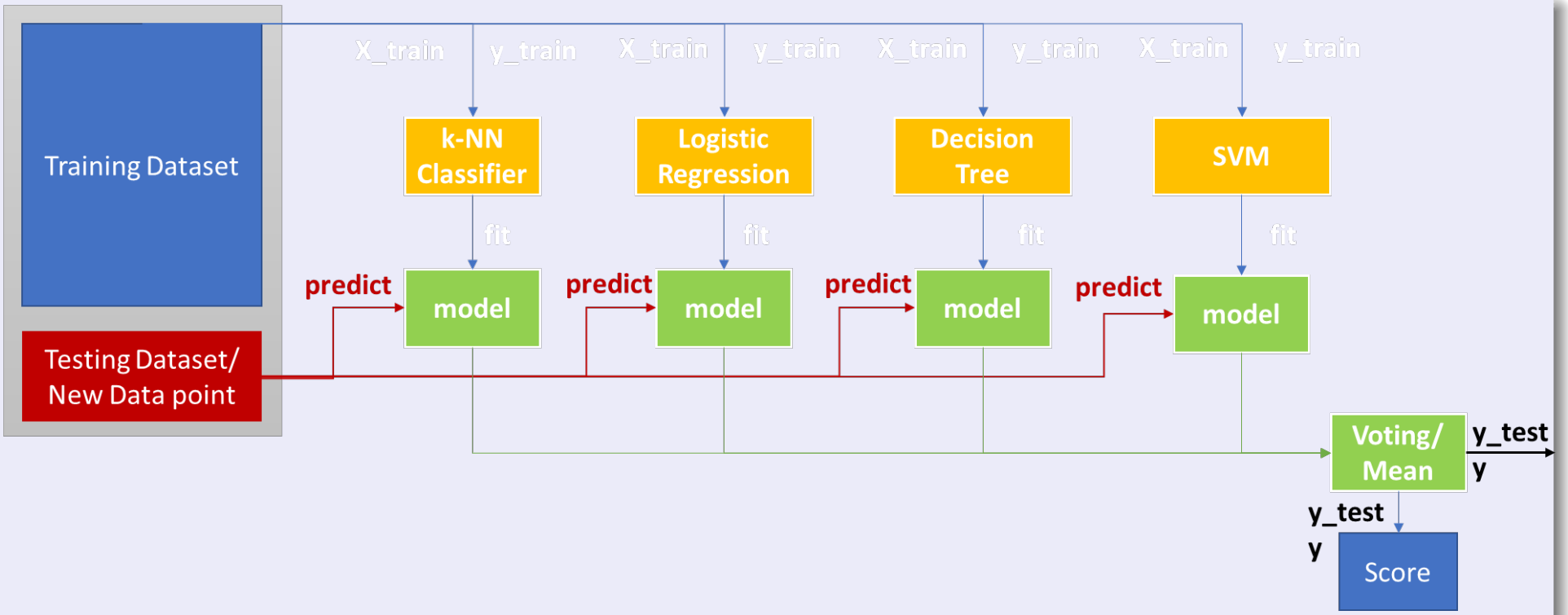
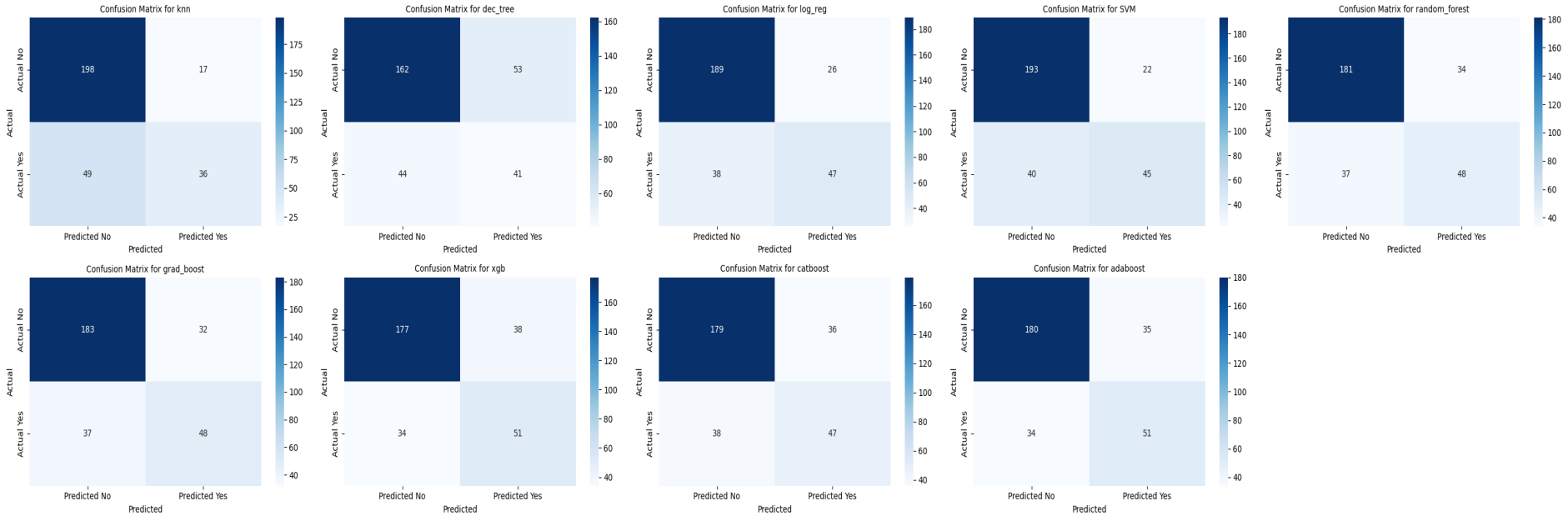


Figure: Ensemble Learning Techniques (Voting/Mean)

Machine Learning Algorithms

- Employed various algorithms for sentiment analysis: **KNN**, **Decision Tree**, **Logistic Regression**, **SVM**, **Random Forest**, **Gradient Boosting**, **XGBoost**, **CatBoost**, **Naive Bayes**, **AdaBoost**.
- Randomly split the dataset into a training set (80%) and testing set (20%).

- **Confusion Matrix (Model Evaluation)**



Automating Machine Learning Pipelines

Models and Accuracy

Models	Accuracy
knn	78%
Decision Tree	69%
Logistic Regression	78.67%
SVM	79.33%
Random Forest	76.67%
Gradient Boosting	77%
XGBoost	76%
CatBoost	75.33%
AdaBoost	77%

- **The ML pipeline of Voting Classifier has a score of:** 76.33%

Why does the ML pipeline of Voting Classifier have 76.33%, smaller than the best Accuracy (SVM with 79.33%)?

Phase 2
Improvement of Automating Machine Learning Pipelines

Improvement of Automating Machine Learning Pipelines

Observation

Most of the algorithms have lower accuracy than the best-performing one.

Question 2

Can the performance be optimized by selecting the top 3 models?

Models and Accuracy (Top 3 Models)

Models	**Accuracy**
knn	78.33%
Logistic Regression	78.33%
SVM	79.33%

- **The ML pipeline of Voting Classifier has a score of:** 79.33%

Comparative Evaluation of Machine Learning Models

Answer 2

With the selection of the top 3 models, the ML pipeline of the Voting Classifier appears to have a better score.

Answer 1

Choosing 8 models but with most of them having low Accuracy, the majority voting appears to make incorrect estimations.

Phase 3
Candidates' Features And Prediction

Candidates' Features And Prediction

A New Data Point

Age
Gender
Education Level
Experience Years
Previous Companies
Distance From Company
Interview Score
Skill Score
Personality Score
Recruitment Strategy

- The prediction is Hired if label $= 1$
- The prediction is Not Hired if laber $= 0$

Conclusions

Conclusions

- **By using the three best models, the accuracy increased by 3%.**
- **The prediction for the new candidate will be made according to the improved version.**

References

- Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019.
- Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- Ravi, S., & Srinivasan, P. *A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges*. IEEE Access, 2020.
- Weiss, A. J. *An Empirical Comparison of Supervised Learning Algorithms for Classification*. Data Mining and Knowledge Discovery, 2019.
- Smith, A., & Doe, J. *Tuning Hyperparameters for Machine Learning Algorithms: A Practical Approach*. Journal of Machine Learning Research, 2018.
- Scikit-learn Documentation. *Scikit-learn*. Available online: `https://scikit-learn.org/stable/`.
- XGBoost Documentation. *XGBoost*. Available online: `https://xgboost.readthedocs.io/`.
- CatBoost Documentation. *CatBoost*. Available online: `https://catboost.ai/docs/`.
- Kaggle Dataset. *Predicting Hiring Decisions in Recruitment Data*. Available online: `https://www.kaggle.com/datasets/rabieelkharoua/predicting-hiring-decisions-in-recruitment-data/code`.

Hyperparameter Tuning Grids

Model	Hyperparameters
K-Nearest Neighbors (KNN)	n_neighbors: [2, 4, 6, 8] p: [1, 2, random.random()]
Decision Tree	criterion: [gini, entropy, log_loss] splitter: [best, random]
Logistic Regression	penalty: [l1, l2, elasticnet, None] solver: [liblinear, saga] multi_class: [auto, ovr, multinomial]
Support Vector Machine (SVM)	kernel: [linear, poly, rbf, sigmoid] decision_function_shape: [ovo, ovr]
Gradient Boosting	n_estimators: [100, 200, 300] learning_rate: [0.01, 0.1, 0.05] max_depth: [3, 4, 5]
XGBoost	n_estimators: [100, 200, 300] learning_rate: [0.01, 0.1, 0.05]
CatBoost	iterations: [100, 200, 300] learning_rate: [0.01, 0.1, 0.05]
AdaBoost	n_estimators: [50, 100, 200] learning_rate: [0.01, 0.1, 0.05]

Automating Machine Learning Pipelines

Hyper-parameter Tuning

Hyper-parameter Tuning refers to setting appropriate parameters for the ML Algorithms that is used during the super- or unsupervised learning procedure.

Ensemble learning

Ensemble learning combine multiple models, each with its strengths and weaknesses, the ensemble can achieve better results than any single model alone.

Automating Machine Learning Pipelines

Pipeline automates all the steps of building and testing an ML model and reduces the required programming effort, not only for individual ML models but also for ensemble learning and hyper-parameter tuning procedures.