Machine Learning Model

Haldun Köktaş

2022-11-06

Transformations

Columns that need different transformations are determined.

Numeric features	One-Hot-Encoded features	Label encoded features
Age	BusinessTravel	DistanceFromHome
MonthlyIncome	Department	Education
NumCompaniesWorked	EducationField	JobLevel
PercentSalaryHike	Gender	StockOptionLevel
TotalWorkingYears	JobRole	JobInvolvement-Performance
TrainingTimesLastYear	MaritalStatus	Satisfaction
YearsAtCompany	Over18	
YearsSinceLastPromotion		
YearsWithCurrManager		

Algorithm Testing

I calculated scores with “Logistic Regression”, “Decision Tres”, “Random Forest” with default parameters and “kNN-Decision Tree” (with 5-fold cross-validation).

The best of them is the “Decision Tree”.

I used “recall” and “precision” as the primary metrics. Because the data is imbalanced and identifying positive labels is more important for the business problem, high recall is a better solution.

I used grid search to find if the search for optimal parameters results in a better model.

Grid search parameters:

Grid Search	criterion	max_depth
Params	[‘gini’, ‘entropy’]	[3, 10, 30]

Comparison of scores for each algorithm tested:

Scores	Logistic Regression (%)	Decision Tree (%)	Random Forest (%)	Random Forest_GridSearch (%)	kNN-Decision Tree (%)
Accuracy	85	97	98	-	96
Precision	62	89	98	88	84
Recall	18	91	87	91	91

While random forest’s precision and accuracy outrun the decision tree, there is a 4% difference regarding the recall.

I selected the decision tree algorithm. It has a higher recall and close precision-recall score (F1).

Feature Importance

The five most important features according to the final model are as follows: (These features might be used for further explorations and possible solutions to the problem)

Feature	Importance (%)
MonthlyIncome	13
TotalWorkingYears	12
PercentSalaryHike	8
Age	7
Satisfaction	7

Testing Phase

I applied transformations to the test set same as the train set. Scores for the test set for "DecisionTreeClassifier() with default parameters":

Score	Test Set (%)
Recall	92
Precision	93
Accuracy	98

Confidence intervals of scores for any given dataset:

Score	Test Set (%)	Population Lower (%	Population Upper (%)
Recall	92	88	97
Precision	93	89	98
Accuracy	98	97	99