Ayda Gizem Kumkumoğlu
Esra Akarçay
Gizem Güneş

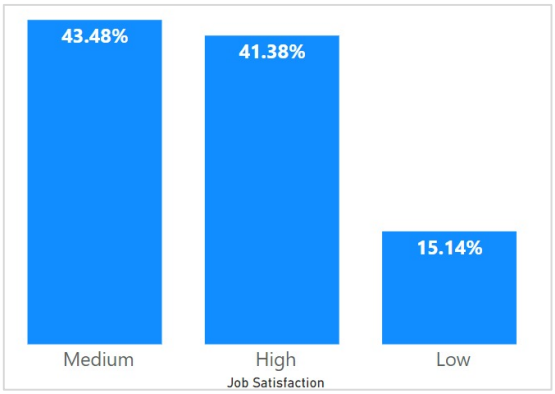
DA 514
HW3-Class Project
19.03.2022

Analysis Report on Kagglers' Job Satisfaction Level and Building a Machine Learning Model

1. Exploratory Data Analysis Findings

The data set about Kagglers has 5529 samples, 53 features and 1 target variable. 2 of 53 features are numerical variables and the rest are categorical. Target variable is named as Job Satisfaction which is 3-class variable including Low, Medium and High satisfaction levels. Project data set is an **imbalanced data set** consist of 43.48% Medium, 41.38% High and 15.14% Low labels.

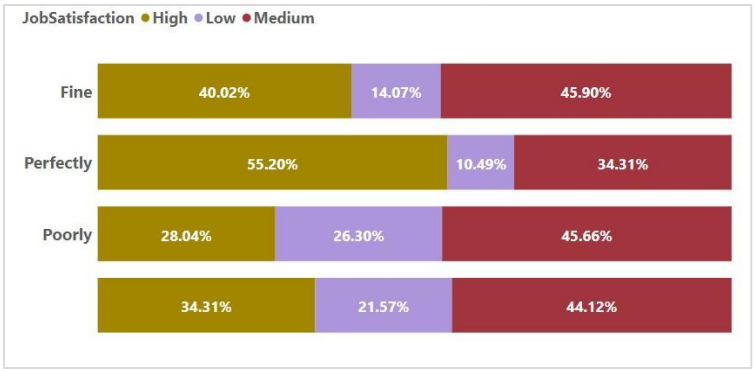
Figure 1: Percentage of Kagglers by Job Satisfaction Level



Kagglers are mostly located in the USA, Asia and West Europe while the rest is located in other regions. Exploratory data analysis findings show that **85% of Kagglers are male and 14% are female**. Kagglers' age has a skewed distribution **with the mean age of 35** approximately. There are suspicious samples in the Age variable namely 0 and 100 and they are ignored during the feature selection process. Female Kagglers have low job satisfaction with 16% which is 2 point greater than Male Kagglers' low job satisfaction ratio. Although, it seems there is no significant difference between Gender and Job Satisfaction level, Gender variable will be included in the model building process.

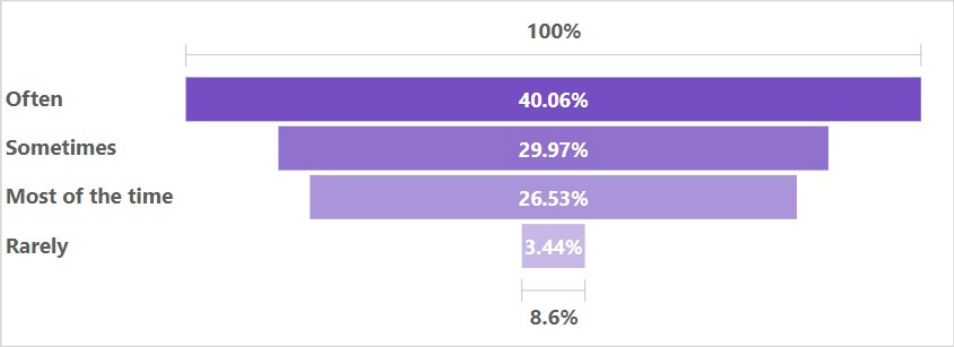
Kagglers mostly have Bachelor Degrees from Computer Science, Mathematics-Statistics and Engineering (Non-computer Focus) majors and their mostly employed in the academic and technology fields. Employment status of Kagglers has 3 categories: employed full-time, employed part-time and independent contractor, freelancer, or self-employed. 85% of the Kagglers are full-time workers. It was observed that Kagglers working as independent contractor, freelancer or self-employed are more satisfied than other categories. Kagglers who are employed full-time, employed part-time and independent ones have high level of job satisfaction ratio of 41%, 38% and 47% respectively.

Figure 2: Title Fit and Job Satisfaction



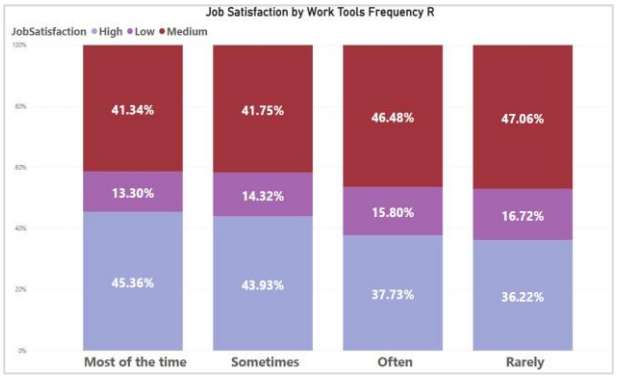
Kagglers specified several current job titles and Data Scientist title is the most frequent one. Also, Kagglers who identified themselves as Data Scientist are dominant in the data set. Kagglers who have claimed that they are perfectly fit to their title are more satisfied (55%) with their jobs than others who claim that they are fine or poorly fit to their title. The data shows that as seniority level of Kagglers increases, their level of job satisfaction increases accordingly.

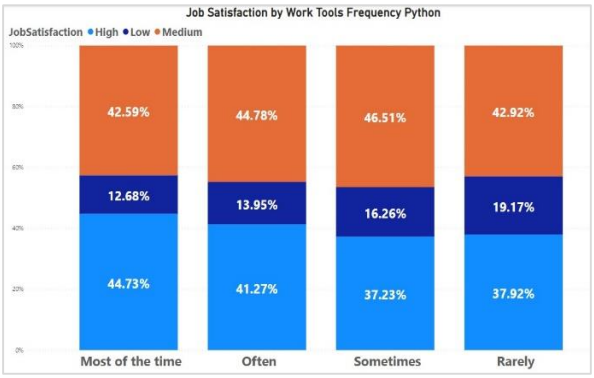
Figure 3: At work how often did you experience these barriers or challenges within the past year?



Exploratory data analysis results show that variety of challenges like company politics, lack of management and financial support for a data science team have a significant impact on the job satisfaction level. Kagglers who face these challenges most of the time in their working environment have the lowest job satisfaction level. It seems that a peaceful, supportive and fair company attitude is important for job satisfaction level. Also, it is an expected finding that Kagglers says they are more satisfied with their jobs when they work remotely. Kagglers who always work remotely have higher job satisfaction.

Figure 4: Job Satisfaction Level of Kagglers by Their Work Tools Frequency





Kagglers' job satisfaction level is dependent on the frequency of tools, platforms and machine learning methods that they use. Python, R and SQL users indicate higher satisfaction levels if they use these languages most of the time or often. Kagglers search for job purposes on different platforms like Kaggle, Blogs, YouTube, Stack Overflow etc., this might be a significant signal for their job satisfaction level since this basically shows their interest on their work. Also, working frequency with different machine learning algorithms has an effect on Kagglers' job satisfaction level. If they work on cross-validation, data visualization, decision trees, logistic regression, neural networks, PCA, random forests and time series analysis most frequently, their job satisfaction level increases since their competence in data science contributes to their job satisfaction level.

2. Preprocessing

The data set includes 53 features, however, 10 of them are unneeded for the model building processes namely ID, CodeWriter, CurrentEmployerType, MLToolsNextYearSelect, MLMethodsNextYearSelect, LanguageRecommendationSelect, PastJobTitleSelect, MLSkillsSelect, MLTechniquesSelect, WorkAlgorithmsSelect. Some of them are like extended version of other features. For example, WorkAlgorithmsSelect includes many machine learning algorithms that Kagglers use; on the other side there are 8 features for different machine learning methods that Kagglers' rank for their use of frequency. We decide that these 8 different features are individually important for prediction and can be easily implemented to the model building while WorkAlgorithmsSelect increases cardinality of the model. Therefore, these 10 unneeded features are dropped from the data set in the preprocessing.

The data set also includes survey type features that measure Kagglers' responses to a specific question. Our approach for these features is to map them as ordinal encoded to analyze the weight of answers. Ordinal encoded features are listed below:

- TitleFit
- LearningPlatformUsefulnessBlogs
- LearningPlatformUsefulnessKaggle
- LearningPlatformUsefulnessCourses
- LearningPlatformUsefulnessProjects
- LearningPlatformUsefulnessSO
- LearningPlatformUsefulnessTextbook
- LearningPlatformUsefulnessYouTube

- WorkProductionFrequency
- WorkToolsFrequencyPython
- WorkToolsFrequencyR
- WorkToolsFrequencySQL

- FormalEducation
- Tenure
- Remote Work

- WorkMethodsFrequencyCross-Validation
- WorkMethodsFrequencyDataVisualization
- WorkMethodsFrequencyDecisionTrees
- WorkMethodsFrequencyLogisticRegression
- WorkMethodsFrequencyNeuralNetworks
- WorkMethodsFrequencyPCA
- WorkMethodsFrequencyRandomForests
- WorkMethodsFrequencyTimeSeriesAnalysis

- WorkChallengeFrequencyPolitics
- WorkChallengeFrequencyUnusedResults
- WorkChallengeFrequencyDirtyData
- WorkChallengeFrequencyExplaining
- WorkChallengeFrequencyTalent
- WorkChallengeFrequencyClarity
- WorkChallengeFrequencyDataAccess
- WorkInternalVsExternalTools

Kagglers' data set includes two different features indicating China, we categorized two of them as China. Also, Country feature has many countries, in order to infer from these countries, we generated a new feature and categorized these countries according to their regions as Region. GenderSelect feature has male, female and other categories that have less samples, therefore, we categorized other samples as Other. One-hot encoded features are listed below:

- GenderSelect
- EmploymentStatus
- DataScienceIdentitySelect
- EmployerIndustry
- Region

- Country
- CurrentJobTitleSelect
- MajorSelect
- WorkMLTeamSeatSelect

Kagglers' data set have null values and we apply different imputation methods in the Pipeline process during model building. Features that have null values are listed below:

ID 0
GenderSelect 10
Country 16
Age 68
EmploymentStatus 0
CodeWriter 0
CurrentJobTitleSelect 2
TitleFit 102
CurrentEmployerType 71
MLToolNextYearSelect 231
MLMethodNextYearSelect 277
LanguageRecommendationSelect 195
LearningPlatformUsefulnessBlogs 2998
LearningPlatformUsefulnessKaggle 2361
LearningPlatformUsefulnessCourses 2588
LearningPlatformUsefulnessProjects 3030
LearningPlatformUsefulnessSO 2533
LearningPlatformUsefulnessTextbook 3327
LearningPlatformUsefulnessYouTube 3116
DataScienceIdentitySelect 1545
FormalEducation 7
MajorSelect 519
Tenure 14
PastJobTitlesSelect 205
MLSkillsSelect 273
MLTechniquesSelect 311
EmployerIndustry 12
WorkToolsFrequencyPython 1282
WorkToolsFrequencyR 2172
WorkToolsFrequencySQL 2529
WorkMethodsFrequencyCross-Validation 2727
WorkMethodsFrequencyDataVisualization 1909
WorkMethodsFrequencyDecisionTrees 2875
WorkMethodsFrequencyLogisticRegression 2403
WorkMethodsFrequencyNeuralNetworks 3562
WorkMethodsFrequencyPCA 3483
WorkMethodsFrequencyRandomForests 3027
WorkMethodsFrequencyTimeSeriesAnalysis 3235
WorkChallengeFrequencyPolitics 3437
WorkChallengeFrequencyUnusedResults 4169
WorkChallengeFrequencyDirtyData 2760
WorkChallengeFrequencyExplaining 4279
WorkChallengeFrequencyTalent 3163
WorkChallengeFrequencyClarity 3805
WorkChallengeFrequencyDataAccess 3809
CompensationScore 1156
WorkDataVisualizations 29
WorkInternalVsExternalTools 116
WorkMLTeamSeatSelect 162
RemoteWork 582
JobSatisfaction 0
EmployerSize 581
WorkProductionFrequency 626
WorkAlgorithmsSelect 426

3. Model Building and Model Selection

In this part of the project, we have been built **11 models** based on different approaches including Logistic Regression, Random Forrest, Gradient Boosting Classification, XGBoost algorithms.

As this data set consists of imbalanced multiclass labels, it was expected to have better F1 scores with algorithms including weighted class distribution. In order to select best parameters, parameter grid was applied.

Table 1: Models and F1 scores

method	TP	FP	TN	FN	f1
Baseline LR	104	55	460	67	0.501698
GB Classifier	0	0	380	58	0.471919
GB Classifier with new parameters	17	8	420	62	0.496330
XGBoost	15	8	424	61	0.505600
XGBoost with new parameters	29	14	420	61	0.508666
Random Forest	102	50	440	55	0.524855
RF with DecisionTreeRegressor iterative imputer	91	44	426	58	0.533324
RF with ExtraTreesRegressor iterative imputer	92	42	442	59	0.524585
RF with KNeighborsRegressor iterative imputer	98	52	435	62	0.530858
RF with BayesianRidge iterative imputer	88	44	420	60	0.529858
Random forest with SMOTENC	69	54	409	64	0.502873

Base model has been built by using Logistic Regression including "balanced" class weight parameter which was resulted in very expensive approach in terms of execution time which lasted for 39 hours yet it is not very successful in terms of F1 score. For the rest of the models, **Randomized Search** has been chosen to find best parameters of the model to be able to reduce fitting time.

Ensemble models combine the decisions from multiple (weak) learners to create a stronger predictive power and stability, therefore, GB and XGBoost algorithms have been built for better classification performance.

The base **GB classifier** model did not perform as well as baseline **LR model**. Therefore, we tuned the parameters and built another GB based model. Unfortunately, parameter tuning did not sufficiently improve the performance and we decided to move forward with another tree-based ensemble algorithm. Same approach has been applied for **XGBoost** as well. However, the F1 score of base and parameter-tuned models were both not satisfactory.

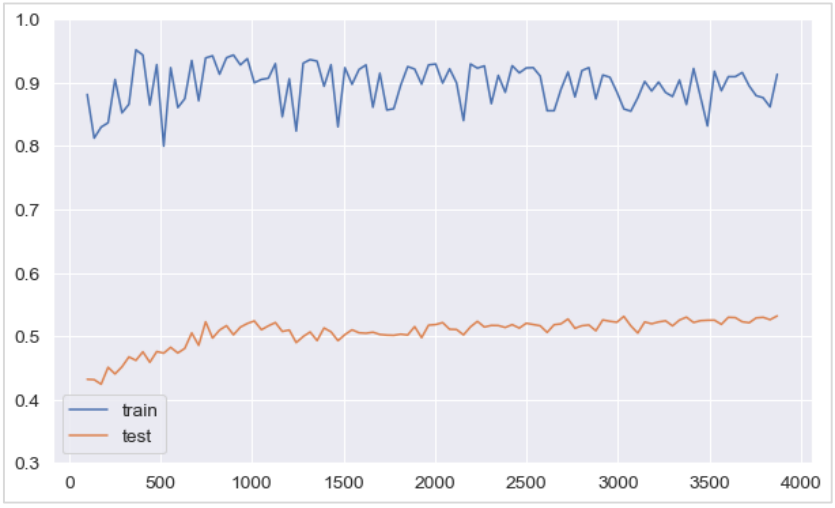
In the next phase, we started to tune **Random Forest** algorithm by checking estimator numbers and each hyper parameters' performance with respect to cross validation score.

After the best hyper parameter range has been found for Random Forest model, final parameter grid have been constructed. The F1 score performance of RF model was better compared to other algorithms. The reasons behind good performance might be the contribution of **fine-tuned hyper parameters and the addition of balanced class weight parameter**.

As final step, SMOTENC technique has been used in order to handle imbalanced classes with both numeric and categorical features. However, the performance of SMOTENC was disappointing. It can be concluded that, for this particular Kagglers data set the fine-tuned and balanced class weight RF model was performed better than other approaches.

Since the Random Forrest model was performed best among others, the improvement actions have been held on this algorithm with different approaches. Four version of iterative imputer have been implemented separately with different algorithms. The best performing model was found as RF with Decision Tree Regressor Iterative Imputer with 53.3% F1 score.

Figure 5: Learning Curve for RF model with Decision Tree Regressor Iterative Imputer



It is observed that there is overfitting with this model since there is a gap between train and test trend. However, we may say that they can converge closer if more data is added.

Table 2: Classification Repot of the Final Model

	Precision	Recall	F1-Score	Support
High	0.57	0.62	0.59	687
Low	0.41	0.36	0.39	251
Medium	0.54	0.52	0.53	721
accuracy			0.54	1659
macro avg	0.51	0.50	0.50	1659
weighted avg	0.53	0.54	0.53	1659

When the class related F1-score is checked, it was found that Low class has the worst performance meaning that the model has poorly performed in classification of Low class compared to High and Medium classes. The most important features based on the model were as following, some of them were investigated in the EDA part.

Age: 0.05714
CompensationScore: 0.04168
TitleFit: 0.03252
LearningPlatformUsefulnessBlogs: 0.01466
LearningPlatformUsefulnessKaggle: 0.01763
LearningPlatformUsefulnessCourses: 0.01733
LearningPlatformUsefulnessProjects: 0.01585
LearningPlatformUsefulnessSO: 0.01704
LearningPlatformUsefulnessTextbook: 0.01424
LearningPlatformUsefulnessYouTube: 0.01546
FormalEducation: 0.01610
Tenure: 0.02295
EmployerSize: 0.03353
WorkProductionFrequency: 0.02740
WorkToolsFrequencyPython: 0.02207
WorkChallengeFrequencyDataAccess: 0.01150
WorkDataVisualizations: 0.03195
WorkInternalVsExternalTools: 0.02184
RemoteWork: 0.03044
CurrentJobTitleSelect_Data Scientist: 0.01340
WorkToolsFrequencyR: 0.02011
WorkToolsFrequencySQL: 0.01991
WorkMethodsFrequencyCross-Validation: 0.02655
WorkMethodsFrequencyDataVisualization: 0.01887
WorkMethodsFrequencyDecisionTrees: 0.01567
WorkMethodsFrequencyLogisticRegression: 0.01706
WorkMethodsFrequencyNeuralNetworks: 0.01584
WorkMethodsFrequencyPCA: 0.01112
WorkMethodsFrequencyRandomForests: 0.01583
WorkMethodsFrequencyTimeSeriesAnalysis: 0.01517
WorkChallengeFrequencyPolitics: 0.08549
WorkChallengeFrequencyUnusedResults: 0.01483
WorkChallengeFrequencyDirtyData: 0.02049
WorkChallengeFrequencyTalent: 0.03923
WorkChallengeFrequencyClarity: 0.01299
DataScienceIdentitySelect_Yes: 0.01491
Region_Asia: 0.01133