I. Definition

Project Overview

Employee attrition is a huge and costly problem for most of the companies. The cost for replacing an employee with new employee involves huge cost [1] [2].

Recent studies shows companies spends huge amount of money to replace an employee .The huge cost may occur due to below reasons

- Training cost
- Hiring cost
- Loss of productivity since new employee takes time to get accustomed to new role etc.,

Problem Statement

Employer invests a huge amount of money in finding a new employee in place of the employee who left the organisation. Since the cost of replacing employees for most employers is very high, if we predict the likelihood of an employee leaving the company, this will lead to actions to improve employee retention as well as possibly planning new hiring in advance. By using the predictive model, we will predict likelihood of an active employee leaving the company it would save a lot of time and money to the employers.

Metrics

Below evaluation metrics [3] is used to evaluate the model performance
Since the dataset is imbalanced, ROC-AUC will be the best evaluation metric to measure the performance of the model

- **Accuracy**: Overall, how often is the classifier correct
- **False positives (FP):** We predicted yes, but actual is No. (Type I error)
- **False negatives (FN)**: We predicted No, but actual is Yes (Type II error)
- **Precision**: True Positives/(True Positives False Positive)
- **Recall**: True Positives/(True Positives False Negatives)
- **ROC-AUC**: Area under the curve

II. Analysis

Data Exploration

The dataset used is an open source data set created by IBM which contains the data about 5000 employees. Since the number of employees leaving will be less when compared to number of active employees, the data set is imbalanced with an attrition event rate of 16%.

The dataset contains several numerical and categorical features providing various information on employee's details (not sensitive) and employment details. The target variable is 'Attrition' which is a binary variable, 0 (active employee), 1 (former employee) .some of the categorical features include Department, Education level, Education Field, Job Role etc. Some of the continuous features includes Age, Number of companies worked, Hike percentage, performance rating etc.

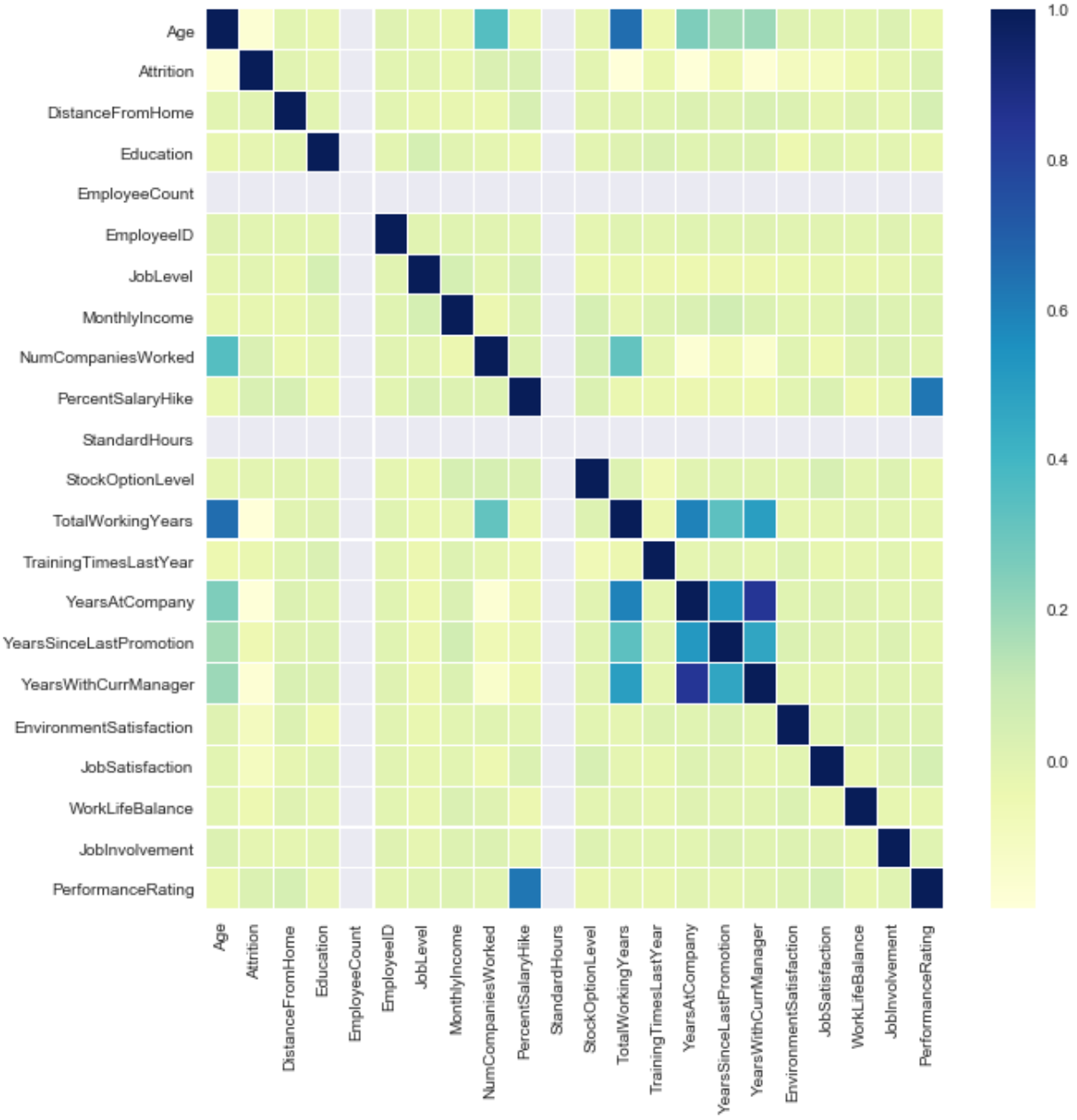
The data dictionary for the employee attrition dataset is:

Variable	Explanation	Levels
Age	Age of the employee	
Attrition(Target Variable)	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees travelled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kilometres	
Education	Education Level	1 'Below College'
		2 'College'
		3 'Bachelor'
		4 'Master'
		5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'

JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	
JobSatisfaction	Job Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low'
		2 'Good'
		3 'Excellent'
		4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock option level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad'
		2 'Good'
		3 'Better'
		4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

Exploratory Visualization

Fig1: A plot showing Correlation of variable sin the employee churn data set



The strongest negative correlations with the Employee Attrition are:

- Total Working Years
- Job Level
- Years InCurrent Role
- Monthly Income

Fig 2: Single employees show the largest proportion of leavers, compared to Married and Divorced.

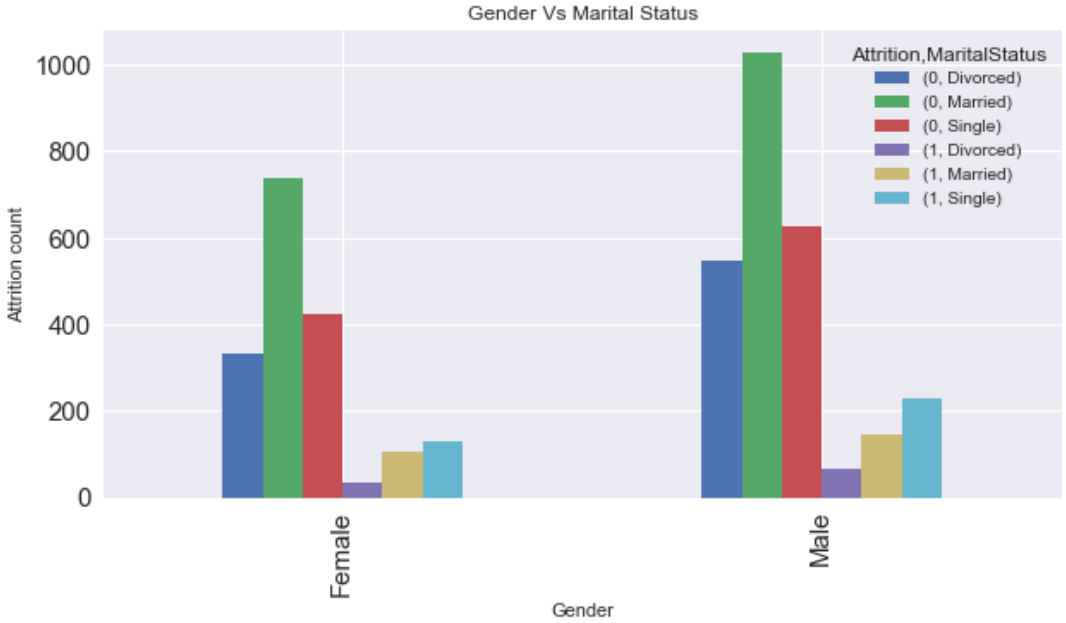


Fig 3: People who travel frequently show higher proportion of leavers when compared to active members

Employee who work as Sales Representatives show a significant percentage of Leavers.

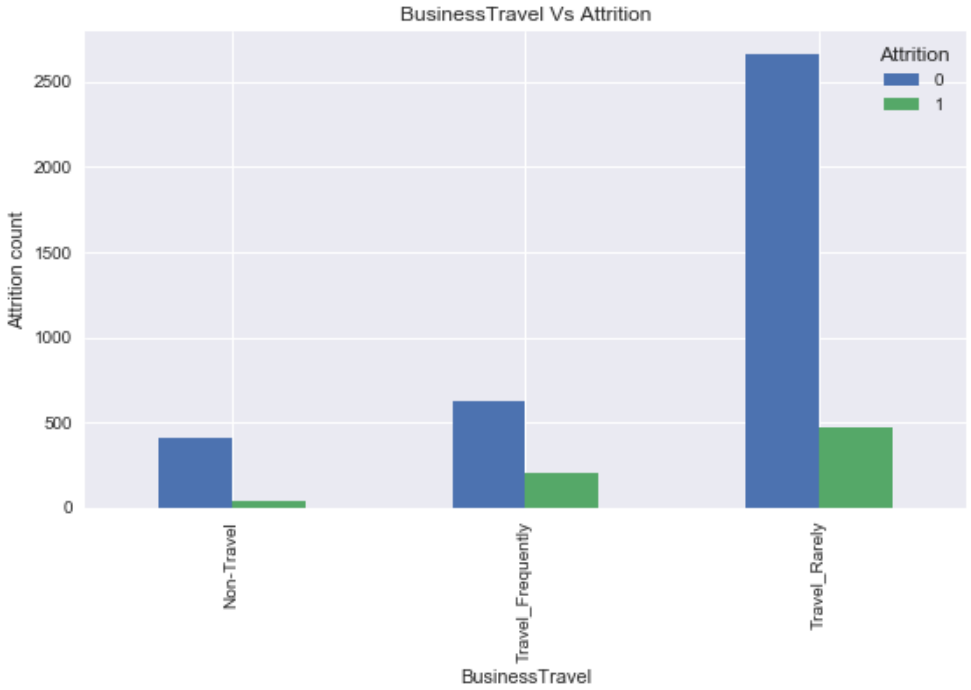
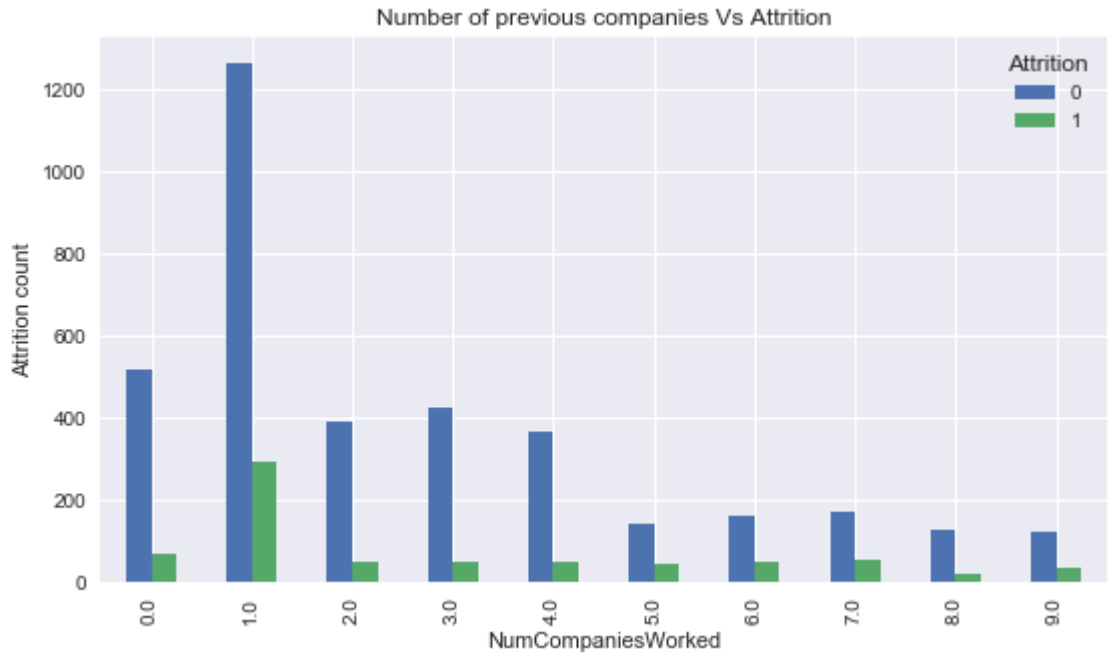


Fig 3: Employees that have already worked at several companies previously show higher proportion of leavers when compared to active members.



Algorithms and Techniques

The employee attrition detection is addressed analytically through below classification models.

- Logistic Regression
- Random Forest
- Support Vector Machines
- KNN-K nearest Neighbours etc.

Based on the metrics such as AUC ROC, Accuracy, precision and Recall Random forest is performing well for this problem.

Random Forest:

A random forest is a Meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Parameters of Random Forest:

n_estimators: integer, optional (default=10) the number of trees in the forest.

Criterion : string, optional (default="gini") the function to measure the quality of split. Supported criteria are "gini" for the Gini impurity and "entropy" for the Information gain. Note: this parameter is tree-specific.

max_depth : integer or None, optional (default=None).Maximum depth of the

Tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split: int, float, optional (default=2) the minimum number of samples required to split an internal node:

min_samples_leaf: int, float, optional (default=1) the minimum number of samples required To be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left And right branches.

Benchmark

Since the problem is a standard **supervised classification problem**, Logistic regression can be used as a starting step. The minimum accuracy would be 80% and minimum ACU-ROC will be 0.5 .The above benchmark metrics is used in comparing the results of the final solution.

III. Methodology

Data Pre-processing

The pre-processing done in the following steps

1. **Missing value treatment:** The missing values are of the particular column is replaced with mean value of that column
2. **Standardization:**
    a. since the age , monthly income columns will be in different scale ,the values are standardized using standard scalar
    b. Standardize features by removing the mean and scaling to unit variance
    c. The standard score of a sample x is calculated as:

        z = (x - u) / s

        Where u is the mean of the training samples or zero if with mean=False, and s is the standard deviation of the training samples or one if with_std=False.

Implementation

The implementation process has two phases

- Training Phase
- Testing Phase

Training Phase:
During this phase, the classifier was trained on the pre-processed training data.

1. Load both the training and testing into memory, pre-processing them as described in Previous section
2. Define the algorithm and training parameters
3. Define the validation Metrics
4. Train the algorithm on the training data

Testing Phase:
The testing data is used to predict the attrition and evaluated with various metrics such as accuracy, precision, Recall and AUC ROC

Refinement

- The hyper parameters[4] are tuned for random forest algorithm by using Grid Search Algorithm .Best parameters obtained in the grid search are used to retrain the training data and model evaluated is the testing data
- By using the grid search parameters F1 score increased around 7 % which is a significant increase.
- The final model has below metrics

Metrics	Accuracy	Precision	Recall	F1 score	AUC -ROC
Value	93 %	75 %	85%	80 %	0.95

IV. Results

Model Evaluation and Validation

- During development, a validation set was used to evaluate the model.
- The final model and hyper parameters were chosen because they performed the best among the tried combinations.
- To verify the robustness of the final model, a K fold cross validation was done and mean metrics were calculated, the calculated mean metrics are in consisted with test metrics.

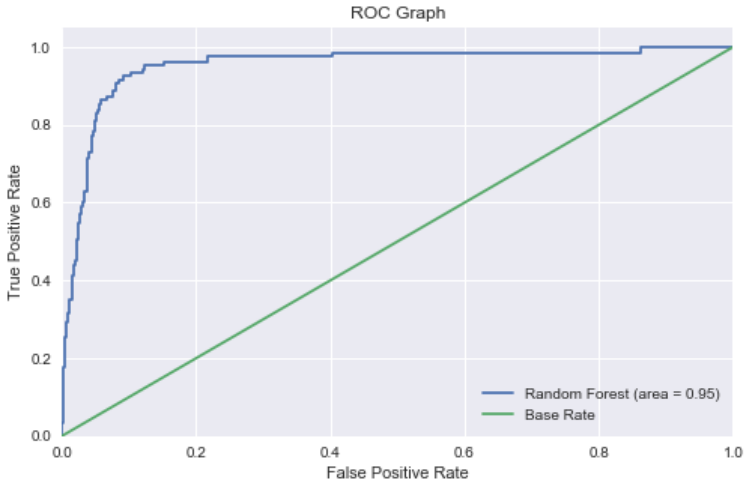
Final Metrics:

Below are the final metrics to evaluate the model

Metrics	Accuracy	Precision	Recall	F1 score	AUC -ROC
Value	93 %	75 %	85%	80 %	0.95

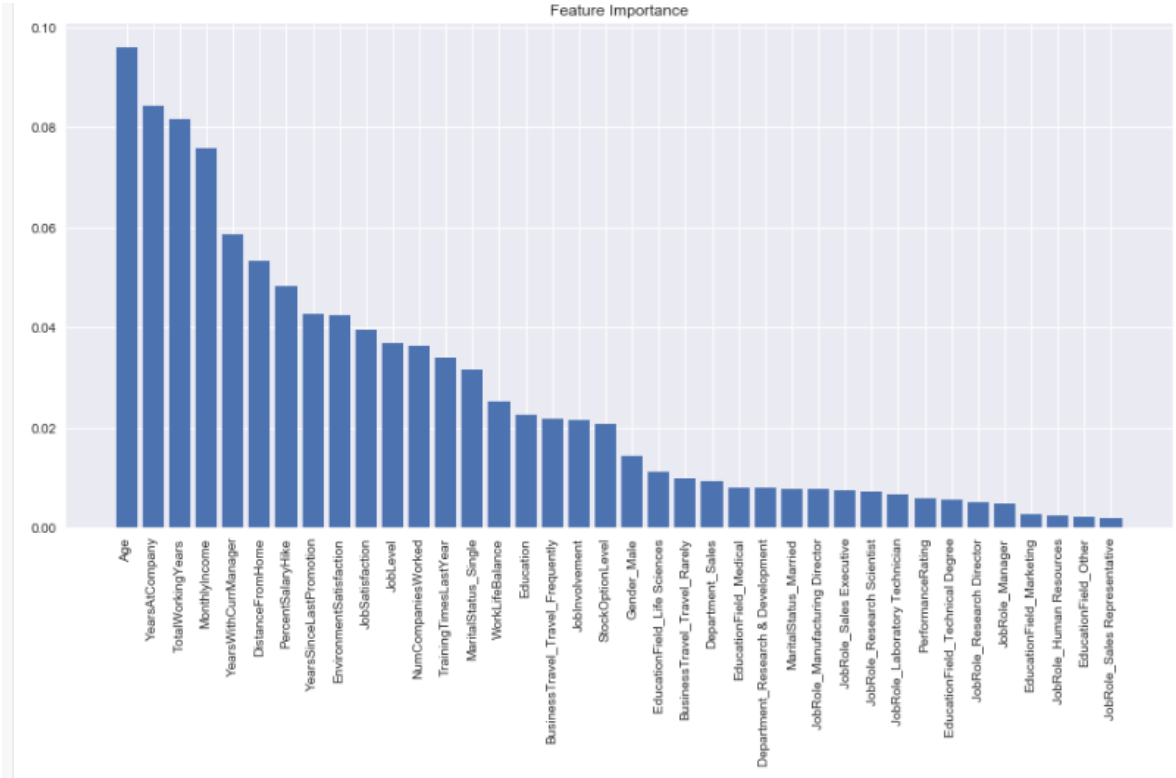
ROC Curve:

ROC AUC value is 0.95 which represents a good model



Feature Importance:

Below is the feature importance of the random forest algorithm



Justification

- ❖ To verify the robustness of the final model, a K fold cross validation [5] was done and mean metrics were calculated, the calculated mean metrics are in consisted with test metrics.
- ❖ Hence the model is performing well on all k fold datasets

V. Conclusion

Reflection

The process used for this project can be summarized using the following steps:

- An initial problem and relevant, public datasets were found.
- The data was downloaded and pre-processed.
- A benchmark was created for the classifier.
- Exploratory data analysis is found to understand the significant variables which affect the target variable
- The classifier was trained using the data (multiple times, until a good set of parameters were found)
- Evaluated the model using evaluation metrics.
- Tested the robustness of the model using k fold cross validation

Improvement

- Some of the other variables such has Login time and logout time, average monthly logging hours, Number of projects worked etc. if available, the model would be still robust.
- As the company generates more data on its employees the algorithm can be re-trained using the additional data and theoretically generate more accurate predictions to identify high-risk employees.
- Employees can be assigning a "Risk Category" based on the predicted probability label such that:
    - **Low-risk** for employees with label < 0.6
    - **Medium-risk** for employees with label between 0.6 and 0.8
    - **High-risk** for employees with label > 0.8

References

[1] https://hal.archives-ouvertes.fr/hal-01556746/document

[2] https://thesai.org/Downloads/IJARAI/Volume5No9/Paper_4-Prediction_of_Employee_Turnover_in_Organizations.pdf

[3] https://www.kdnuggets.com/2018/04/right-metric-evaluating-machine-learning-models-1.html
[4] https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e
[5] https://machinelearningmastery.com/k-fold-cross-validation/