**Human Resources Analytics**

**TASK1:**

Dataset description:

The dataset has 10,500 rows and 10 features. It has description about the employees in a firm (Satisfaction level of employees, Employees last evaluation, number of projects, average monthly hours of the employee, no of years an employee has been employed in the company, no of work accidents, promotion, department, salary). These are the independent variables.

The dependent variable/target is “left” for which left =1 indicates the employee left the company and left=0 indicates that the employee didn’t leave the company.

Business understanding:

The data set concerns a big company that wants to understand why some of their best and most experienced employees are leaving prematurely. The company also wishes to predict which valuable employees will leave next.

Reason to choose the dataset:

The dataset is very interesting since this business case applies to every company irrespective of company’s technology, domain and location of the company. It’s important to understand employee attrition patterns and make strategic decisions to retain good employees. The dataset seems to be comparatively bigger than the previous one and it would be interesting to understand the behavior of the predictive models on a dataset like this

Understanding Variables:

The dataset is a 10-column data set having 15000 rows.

|  |  |
| --- | --- |
| Variable Name | Description |
| satisfaction\_level | Highest being 1 and lowest is 0.09. And there are 10187 employees with satisfaction above 0.5 |
| number\_project | There are employees who are assigned up to 7 project and as least as 2 projects. |
| average\_montly\_hours | On an average, employees spend 200 hours/month in office. |
| time\_spend\_company | The company has employees whose stay varied from 2 to 10 years |
| left | This is the variable of interest. Out of 15K in dataset around 11K employees has not left the company |
| promotion\_last\_5years | Only around 300 employees are promoted in last 5 years |
| sales | There are around 6 major departments. And maximum(6K) belongs to sales department |
| salary | | classified into high/medium/low. And maximum employees belong to low(7K) |

This table describe the characteristics of each features. It shows the different statistical measures of central tendency and variation. The attrition rate is equal to 24%, the satisfaction level is around 62% and the performance average is around 71%. We see that on average people work on 3 to 4 projects a year and about 200 hours per months.

Analytic solution :

There are two goals: first, we want to understand why valuable employees leave, and second, to predict who will leave next.

Initial Findings:

1. Numerous plots shown below are drawn to understand the dataset further.

|  |  |
| --- | --- |
| Screen%20Shot%202017-10-17%20at%2010.11.54%20AM.png | **Co-relation Plot:**  Correlation plot is drawn to check the correlation between the features and the target variable “left”. As per the plot shown, the most 2 co-related features are Age followed by Estimated Salary  The size of the bubbles reveals the significance of the correlation, while the color presents the direction.  . |
| Screen%20Shot%202017-10-17%20at%2010.17.05%20AM.png | **Distribution of Purchased values in the dataset:**  The variable can have 0 (Didn’t leave) and 1 (left). From the plot, it can be seen that there are around 30% of the people that dis nit leave the company and 60% of them left. |
| Screen%20Shot%202017-10-17%20at%202.33.41%20PM.png  Screen%20Shot%202017-10-17%20at%202.33.34%20PM.png | The monthly average hours working, last evaluation, satisfaction level and salary of people who left the company is shown in the graphs below.  We can see why we don't want to retain everybody. Some people don't work well as we can see from their evaluation, but clearly there are also many good workers that leave.  In the total of 15 000 employees that compose our database, the number people that have left are 3571.  The total of employees that received an evaluation above average, or spend at least four years in the company, or were working on more than 5 projects at the same time and still have left the company are 2014.These are the people the company should have retained |
| Screen%20Shot%202017-10-17%20at%2011.02.36%20AM.png | The correlation plot describes why the above described good people leave the company.  On average valuable employees that leave are not satisfied, work on many projects, spend many hours in the company each month and aren't promoted. |

1. Below plot is shown below to understand the dataset further.

|  |  |
| --- | --- |
| ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.17.52%20 | **Variable Importance:**  This plot is drawn to check what is the importance of each variable on the dependent variable “left”. From the plot, it can be seen that last\_evaluation, salary and satisfaction\_level are top 3 features that are affecting the output the most. |

1. Feature Scaling:

Feature scaling is performed on the variables of this dataset to make sure they are all in the same range and it also helps with data visualization/manipulation and to compare performance with other models.

1. Dataset Split:

Dataset is split into Training and Testing set in 70:30 proportion in KNN and 70-15-15 in ANN.

1. Cross-Validation:

**K-NN:** Cross-Validation is performed using knn.cv method in K-NN. Cross validation is performed implicitly with this method where the training set is split appropriately.

**ANN:** Cross-Validation is performed by using below.

**validation\_frame = as.h2o(validation\_set),**

**nfolds =10,**

**keep\_cross\_validation\_predictions=TRUE,**

1. Variable Importance:

Variable importance is performed using h2o.varimp method.

**Modeling:**

Now we want to predict which valuable employee will leave next.

**TASK2:**

SVM Model (Using kernels):

SVM model is trained on several kernels – Linear, Radial, Polynomial and Sigmoid. The accuracies for each is given in the table below.

In polynomial kernel, degree and gamma were given different values and accuracy was checked against each and is chosen as 3 and gamma = 1 since that yielded better accuracy.

In Sigmoid kernel, gamma was fixed on 1 after substituting different values.

TASK3: Decision Tree:

The decision tree is split based on gini calculation. Below is the decision tree before and after pruning.

The model seems to be already generalized with the decision tree and hence pruning did not yield any extra benefit on the decision tree and provides the same tree and same accuracy. Pruning is tried on different values of cp and cp=0.025 was chosen since that is the cp with lowest error.

|  |  |
| --- | --- |
| Decision Tree split – Before Pruning | Decision Tree – After pruning |
| Screen%20Shot%202017-10-17%20at%2012.47.40%20PM.png | ../../../../../../Desktop/Screen%20Shot%202017-10-17%20at%203.00.40%20 |

**TASK 4:**

Boost Model:

The dataset is trained on Xgboost boosting algorithm. Prediction is performed on the testing set and gained an accuracy of 85%.

Pruning is applied on Xgboost by adding gamma parameter to the xgboost package. Accuracy is tested against different values of gamma and gamma=0.7 is chosen to achieve best accuracy

**TASK 5:**

**K-NN:**

Split the data into training and testing with 70/30 proportion. K-NN is executed with K=3 initially that gave accuracy of 95%. We then run the loop to see the accuracy for different values of k. The model then returned highest accuracy for K=1. The graph is shown in the table below.

Cross validation is performed for K-NN using knn.cv method.

**TASK 6:**

**ANN:**

We have used h2o to implement ANN. The data is split into 70-15-15. Cross validation is performed using validation frame parameter in the h2o.deep learning method.

We have changed different parameters using grid search. Different values for activation ("Rectifier", "RectifierWithDropout", "Maxout", "MaxoutWithDropout","Tanh","TanhWithDropout") and hidden ((50,50), (100,100), (200,200), (300,300)) are used to get the best model.

Once we have the best model, we are then using these parameters to predict on test data using h20.performance.

|  |  |
| --- | --- |
| ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.28.24%20 | 1. Rectifier  2. RectifierWithDropout  3. Maxout  4. MaxoutWithDropout  5. Tanh  6. TanhWithDropout  **As we see from the plot, the activation is high for Maxout (3).** |
| ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.39.20%20 | Hidden Layers = ((10,10),(25,25),(50,50),(100,100),(200,200),(300,300),(350,350) with Activation = Maxout (since Maxout gave us more Accuracy)  Accuracy values are – (0.9717604 ,0.9707480 ,0.9776354, 0.9752724 ,0.9811440 ,0.9795605, 0.9821553 ,0.9823285)  **Accuracy is maximum for 8 which is 98.23% with (100,100)** |
| ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.20.43%20../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.22.39%20 | The first figure shows the duration in min (clock time) that it took to train and validate the input dataset v/s mse (mean squared error).  The second figure shows the duration in min (clock time) that it took to train and validate the input dataset v/s logloss.  Logloss and mse decreases as we train for longer time with more observations. |
| ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.23.22%20../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.21.37%20 | The first figure shows the duration in min (clock time) that it took to train and validate the input dataset v/s AUC.  The second figure shows the duration in min (clock time) that it took to train and validate the input dataset v/s classification error.  AUC increases as the dataset is trained for longer time and classification error decreases. |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| COMPARISION TABLE – HR ANALYTICS | | | | | | | | | | | |
| **Model** | **Method** | **Confusion Matrix** | | **Accuracy** | **ROC Curve** | | | **Sensitivity** | **Specificity** | **Model Remarks** | |
| **Support Vector**  **Machine** | Kernel: linear | Screen%20Shot%202017-10-17%20at%2011.38.52%20AM.png | | 0.7742 | Screen%20Shot%202017-10-17%20at%2011.40.03%20AM.png | | | 0.7991 | 0.5589 | The accuracy of the SVM model when Linear kernel is used is 59.18% which indicates that the model is not so well separable linearly..  The accuracy is low compared to Decision Tree. | |
| Kernel: radial | Screen%20Shot%202017-10-17%20at%2011.59.53%20AM.png | | 0.9522 | Screen%20Shot%202017-10-17%20at%2011.59.59%20AM.png | | | 0.9682 | 0.9007 | The accuracy of the SVM model when Radial kernel is used is 95.22%. This kernel performs better than linear on this data. | |
| Kernel: polynomial, degree =3  Gamma=1 | Screen%20Shot%202017-10-17%20at%2012.05.45%20PM.png | | 0.9571 | Screen%20Shot%202017-10-17%20at%2012.35.23%20PM.png | | | 0.9731 | 0.9065 | Polynomial Kernel is predicting with a good accuracy and higher than linear and radial with 95.7%. | |
| Kernel: sigmoid: gamma=1 | Screen%20Shot%202017-10-17%20at%2012.32.43%20PM.png | | 0.5479 | Screen%20Shot%202017-10-17%20at%2012.32.49%20PM.png | | | 0.70333 | 0.05042 | Sigmoid Kernel is not predicting with a good accuracy and its lower than linear and radial with 54.7%. It is the lowest of all. | |
| **Decision Tree** | Without pruning | Screen%20Shot%202017-10-17%20at%2012.48.00%20PM.png | | 0.9711 | Screen%20Shot%202017-10-17%20at%2012.49.55%20PM.png | | | 0.9739 | 0.9617 | This model is providing accuracy of 97.1%. | |
| After pruning with best cp=0.025 | Screen%20Shot%202017-10-17%20at%2012.59.52%20PM.png | | 0.9711 | Screen%20Shot%202017-10-17%20at%2012.51.25%20PM.png | | | 0.9886 | 0.9150 | After applying pruning, this model is providing an accuracy of 97.1% again. Since the tree is not big, pruning on such trees doesn’t yield varied result. Hence pruning on this dataset is not improving accuracy or ROC. | |
| **Xgboost** | **Xg boost** | **Screen%20Shot%202017-10-17%20at%209.07.57%20AM.png** | | **0.9787** | **Screen%20Shot%202017-10-17%20at%2012.43.46%20PM.png** | | | **0.9763** | **0.9870** | **The boosting model has an accuracy of 97.87% before pruning.** | |
| Changing gamma for pruning: best =0.75 | 1.png | | 0.9798 | Screen%20Shot%202017-10-17%20at%2012.44.31%20PM.png | | | 0.9774 | 0.9880 | The boosting model after pruning gives 97.98% accuracy. The pruning is done here by adjusting gamma values. Gamma = 0.75 is the best gamma for which accuracy is the highest. | |
| **COMPARISON BETWEEN THE MODELS – HR ANALYTICS DATASET** | | | | | | | | | | |
|  | **CONFUSION MATRIX** | | **ACCURACY** | **SENSITIVITY** | | **SPECIFICITY** | **ROC CURVE** | | **NO. OF NEIGHBORS V/S ACCURACY** | |
| K-NN  (With k=3) | ../Desktop/Screen%20Shot%202017-11-06%20at%201.38.04%20PM.png | | 0.9548 | 0.9763 | | 0.8776 | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%207.46.27%20 | | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%205.59.29%20  k=1 has the highest accuracy as expected since this is a binary problem. | |
| K-NN  (With k=1) | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%206.28.57%20 | | 0.9656 | 0.9895 | | 0.8987 | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%207.05.45%20 | | This model is providing accuracy of 95.07. After choosing the optimal k for which accuracy is higher, which is k=1, the accuracy increased to 96.66 | |
| Best ANN after grid search | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.48.49%20 | | 0.967 | 0.9815 | | 0.9200 | ../../../../../../Desktop/Screen%20Shot%202017-11-05%20at%2011.19.05%20  AUC = 0.98527 | | This model has the accuracy of 98.5% with Activation = Maxout and Hidden = (100,100) for which accuracy is maximum. | |

**All Models - Comparison and Rank:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **AUC** | **Accuracy** | **Rank** |
| Decision Tree | 0.9518 | 0.9711 | 4 |
| XgBoost | 0.9594 | 0.9798 | 3 |
| SVM (Polynomial, degree 3) | 0.9423 | 0.9571 | 5 |
| K-NN | 0.9666 | 0.9656 | 2 |
| ANN | 0.98527 | 0.967 | 1 |

AUC is the measure of TPR v/s TNR and this value is a better measure than accuracy since it gives better understanding about the predictions. Though XgBoost has the highest accuracy and lower AUC compared to ANN. ANN has the highest AUC (considering the trade off between AUC and Accuracy) and hence the best model of all.

**Conclusion:**

On average people who leave have a low satisfaction level, they work more and didn't get promoted within the past five years.

1. If you're successful and overworked, you leave.

2. If you're unhappy and overworked, you leave.

3. If you're unhappy and underworked, you leave.

4. If you've been at the company for more than 6.5 years, you're more likely to be happy working longer hours.

**Conclusion:**

After comparing accuracies of all the models, **ANN** is the better model with Area Under Curve = 98.5%. This means that the new data will be predicted better with 96.7% accuracy.

**Additional things to get a better model:**

1. We can tune the other hyper parameters such as epocs, learning rate, train samples per iteration, seed etc in learning methods.
2. We can add more data/observations to generalize the model to predict with better accuracy.

On average people who leave have a low satisfaction level, they work more and didn't get promoted within the past five years.

1. If you're successful and overworked, you leave.

2. If you're unhappy and overworked, you leave.

3. If you're unhappy and underworked, you leave.

4. If you've been at the company for more than 6.5 years, you're more likely to be happy working longer hours.

From business perspective, here’s the conclusion –

Retain the following employees who have a higher probability of leaving –

1. The employees with low satisfaction, too little projects/work cannot handle enough projects, which eventually makes them leave (either by choice or not).
2. Those with low satisfaction and probably work too much, and it seems likely that they quit.