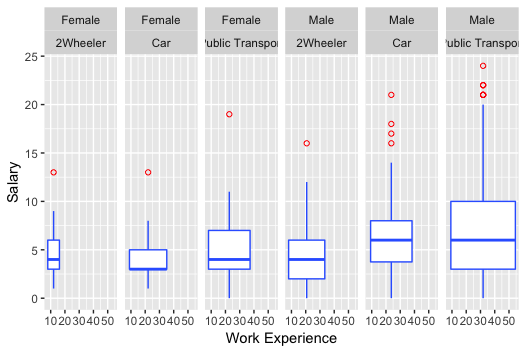
**EDA** :

* The summary stats of our data tell us that there are 444 records in total. Out of which a majority of them are males (316) and the rest are females.
* The age column ranges from 18 to 43, with the mean and median positioned around 27 and 3rd quartile at 30 – implying that 75% our data contains information of people under the age of 30.
* 75% of the people are Engineers.
* 25% of the people have a MBA degree.
* Work experience ranges from 0 to 24, median at 5.0, mean at 6.3 and third quartile at 8.0 – confirming again that our data contains information about relatively young people.
* Salary ranges from 6.50 LPA to 57.00 LPA with an average of 16.24 LPA.
* Distances that the people travel everyday lie between 3.20 to 23.40 kilometers. On an average people in this data travel 22.64 kms (twice the average) a day.
* 23% of the have a license.
* The people who use public transport is the highest, constituting more than 2/3rd of the data. It is followed by 2 wheeler(83) and car(61).
* Also from the MBA column, we observe there is one NULL value.



**Treating NULL values**

As we had observed during EDA, we had one NULL value in the MBA column which needs to be removed. Hence, we will be using KNN function inside VIM package to remove that.

After treating the NULL values, we proceed to building a SVM model.

**SVM MODEL**

We build a Support Vector Machine model using the below code.

svm\_tune <- tune(svm, Transport~., data = Trans\_avail\_Final, ranges = list(cross = 7, epsilon = seq(0,1,0.01), cost = 2^(2:9)))

After building the model, we are looking at the model findings and using that we will be finding the best parameters by tuning cost and epsilon.

From the initial model, we get the below findings.

> print(svm\_tune)

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:

cross epsilon cost

7 0 64

- best performance: 0.2004545

After fine tuning, we get the below best values for Cross, Epsilon and Cost.

> svm\_tune$best.parameters$epsilon

[1] 0

> svm\_tune$best.parameters$cost

[1] 64

> svm\_tune$best.parameters$cross

[1] 7

Now we will deploy the SVM model that can be used to separate the data into multiple hyperplanes. Each hyperplane of the data will be contain data that is homogeneous in nature.

Now we will proceed with K-fold cross validation.

**7-Fold Cross Validation Model using Linear Kernel: Model 1**

svm\_model<-svm(Trans\_avail\_Final$Transport~., data=Trans\_avail\_Final, kernel="linear", tolerance=0.0001, shrinking=TRUE, cross=7, fitted=TRUE)

By running the above code and building a SVM model, we get the below output.

Call:

svm(formula = Trans\_avail\_Final$Transport ~ ., data = Trans\_avail\_Final,

kernel = "linear", tolerance = 1e-04, shrinking = TRUE, cross = 7,

fitted = TRUE)

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

gamma: 0.1111111

Number of Support Vectors: 199

( 95 83 21 )

Number of Classes: 3

Levels:

2Wheeler Car Public Transport

7-fold cross-validation on training data:

***Total Accuracy: 77.47748***

Single Accuracies:

82.53968 73.01587 81.25 74.60317 81.25 71.42857 78.125

And using this model we try to predict the Class of the data and try to understand the accuracy of the model. We are getting a total accuracy of 77.47%. Let us try to build a ***confusion matrix*** from this model.

**Predicted class**

**Actual class** 2Wheeler Car Public Transport

2Wheeler 5 4 74

Car 1 53 7

Public Transport 0 5 295

As an additional step, let us try to implement the same model with the kernel as “Radial” and check the accuracy.

**7-Fold Cross Validation Model using Radial kernel: Model 2**

We will run the below code and will build a model with a radial kernel.

svm\_model<-svm(Trans\_avail\_Final$Transport~., data=Trans\_avail\_Final, kernel="radial", tolerance=0.0001, shrinking=TRUE, cross=7, fitted=TRUE)

summary(svm\_model)

Now we will get the below output.

Call:

svm(formula = Trans\_avail\_Final$Transport ~ ., data = Trans\_avail\_Final,

kernel = "radial", tolerance = 1e-04, shrinking = TRUE, cross = 7,

fitted = TRUE)

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

gamma: 0.07692308

Number of Support Vectors: 179

( 88 67 24 )

Number of Classes: 3

Levels:

2Wheeler Car Public Transport

7-fold cross-validation on training data:

***Total Accuracy: 90.54054***

Single Accuracies:

95.2381 85.71429 90.625 95.2381 81.25 93.65079 92.1875

We could see that the accuracy has improved drastically compared to the previous one. It is at 90.54%. Let us build the confusion matrix.

**Predicted class**

**actualclass**  2Wheeler Car Public Transport

2Wheeler 51 1 31

Car 0 61 0

Public Transport 10 0 290

Let us build a SVM model using best parameters obtained from the optimization steps.

**SVM Model using Best Parameters obtained from Tuning : Model 3**

From the best parameters obtained in the previous steps, we will build an SVM model and we will calculate the accuracy from the confusion matrix.

mysvm <- svm(Trans\_avail\_Final$Transport~., data=Trans\_avail\_Final, cost = svm\_tune$best.parameters$cost, epsilon = svm\_tune$best.parameters$epsilon)

summary(mysvm)

Call:

svm(formula = Trans\_avail\_Final$Transport ~ ., data = Trans\_avail\_Final,

cost = svm\_tune$best.parameters$cost, epsilon = svm\_tune$best.parameters$epsilon)

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 64

gamma: 0.09090909

Number of Support Vectors: 201

( 103 70 28 )

Number of Classes: 3

Levels:

2Wheeler Car Public Transport

We will now summarize the model above, we will try to obtain a ***confusion matrix.***

**Confusion Matrix and Statistics**

**Predicted class**

**Actual class**  2Wheeler Car Public Transport

2Wheeler 52 0 31

Car 0 61 0

Public Transport 10 0 290

**Overall Statistics**

**Accuracy : 0.9077**

95% CI : (0.8768, 0.9329)

No Information Rate : 0.723

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8021

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 2Wheeler Class: Car Class: Public Transport

Sensitivity 0.8387 1.0000 0.9034

Specificity 0.9188 1.0000 0.9187

Pos Pred Value 0.6265 1.0000 0.9667

Neg Pred Value 0.9723 1.0000 0.7847

Prevalence 0.1396 0.1374 0.7230

Detection Rate 0.1171 0.1374 0.6532

Detection Prevalence 0.1869 0.1374 0.6757

Balanced Accuracy 0.8788 1.0000 0.9111

We could observe that the accuracy in predicting the classes has improved for this model compared to the previous model. The accuracy is 90.77% for this model which is the best among all the three models. Hence, we will use this model for test data predictions.

**QUESTION 1 :**

**-Build a model that best explains the employee’s decision to use cars as the main means of transport?**

We will now modify the data such that Transport column says whether or not the person uses ‘Car’ as his/her mode of transport. So we will code 1 for the rows which have the value ‘Car’ and 0 for all the rest. This will help us to understand the factors which will make the users to choose ‘Car’ as their mode of transport.

For this we will build a regression model by taking all the variables into consideration and we will eliminate the variables which do not play a major factor using stepAIC.

Reg\_Trans <- glm(Trans\_avail\_Final\_boost$Transport ~.,family=binomial(link='logit'),data=Trans\_avail\_Final\_boost)

summary(Reg\_Trans)

The summarized results of regression are below.

Call:

glm(formula = Trans\_avail\_Final\_boost$Transport ~ ., family = binomial(link = "logit"),

data = Trans\_avail\_Final\_boost)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.99451 -0.04226 -0.00732 -0.00051 2.27156

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -72.7701 16.0018 -4.548 5.43e-06 \*\*\*

Age 2.2608 0.5263 4.296 1.74e-05 \*\*\*

Gender.Female 1.7067 0.8336 2.047 0.040632 \*

Gender.Male NA NA NA NA

Engineer 0.8573 0.9137 0.938 0.348139

MBA -1.9360 0.9094 -2.129 0.033261 \*

Work.Exp -1.1991 0.3616 -3.316 0.000913 \*\*\*

Salary 0.1853 0.0720 2.573 0.010074 \*

Distance 0.4907 0.1409 3.483 0.000497 \*\*\*

license 2.7089 0.8634 3.137 0.001705 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 355.371 on 443 degrees of freedom

Residual deviance: 63.263 on 435 degrees of freedom

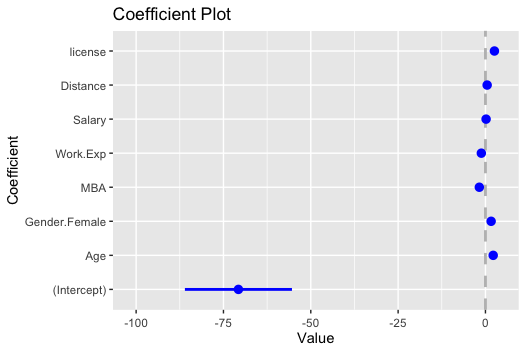
AIC: 81.263

Number of Fisher Scoring iterations: 10

We could see that the factors Age, Work Experience and Distance are highly significant in the user choosing a car for transport. We also have license being a relatively less, still significant factor in the play.

Now we will run StepAIC to determine the most important factors so that we could optimize our regression model for better results.

From the StepAIC process, we obtain the coefficients and then we plot them.



Then we calculate VIF for the model using the below function.

> vif(reg\_transport\_final)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | Female | MBA | **Work.Exp** | Salary | Distance | License |
| 11.474191 | 1.437833 | 1.368433 | 16.645167 | 3.981423 | 1.718773 | 1.731569 |

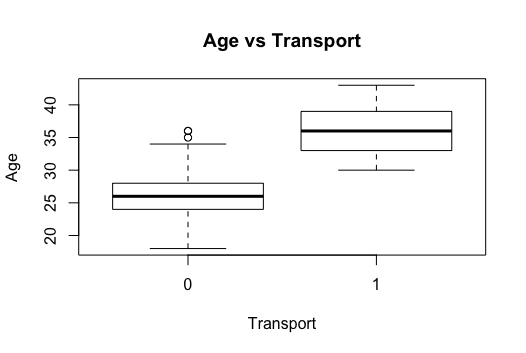
We have obtained the above VIF values. From the values, the ones which have a value greater than 10, they have a strong multicollinearity. We see that Age and Work Experience have values greater 10.

Then we perform Regularization.

<regularization>

**KEY OBSERVATIONS FOR CRITICAL VARIABLES AFTER REGULARISATION**

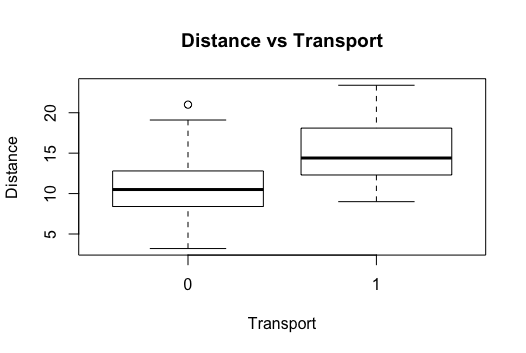
* Users who are using car as transport have Median age much higher than non-car passenger. So higher age seems to be a driving factor for transport mode selection.



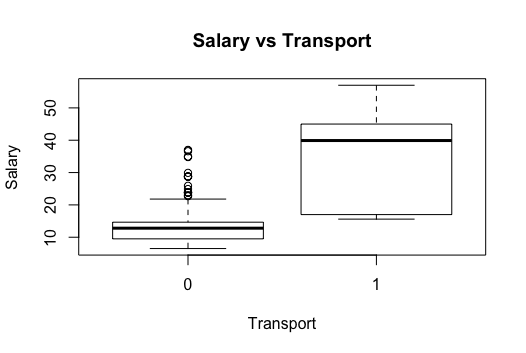
* Users who are using car as transport have Median work experience much higher than non-car passenger. So higher work experience seems to be a driving factor for transport mode selection. This is also clear from age as higher work experience employee will have higher age.



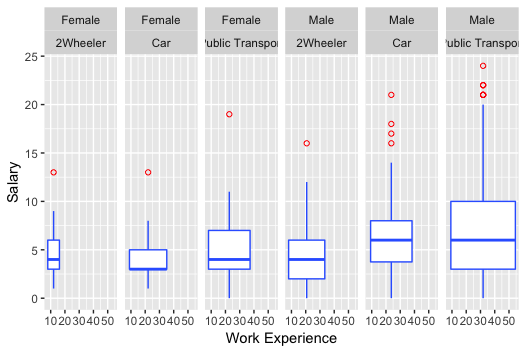
* Users who are travelling long distance prefer to use car as prefered mode of transport



* Higher salary seems to be driving factor for choosing car as preferred one but there are quite a few outliers for non-car owner also who are earning higher salary



* No of Men seems to be much higher w.r.t Women for preferred mode of transport as Car or Non-Car. This may be due to gender-inequality in Job.



**QUESTION 2:**

**-What would your predictions regarding their choice of transport be for the following two employees?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Engineer | MBA | Work Exp | Salary | Distance | license |
| 25 | Male | 0 | 0 | page2image396463362 | 10 | 5 | page2image396440321 |
| 25 | Female | 1 | 0 | 2 | 10 | 5 | 0 |

For this question, we will use the MODEL 3 to predict the classes i.e., the mode of transport for the given data.