Case Study 2

Angelo Bravo

12/6/2019

```
#install.packages("knncat")
library(knncat)
library(caret)
library(dplyr)
library(MASS)


df <- read.csv("/Users/angelobravo/Downloads/MDS-6306-Doing-Data-Science-Fall-2019-Master-7/Unit 14 and


set.seed(123)
#####################attempting to fit knn-model
df1 <- df[complete.cases(df), ]

df1$labels <- ifelse(df1$Attrition == "No", 0, 1)
df1$labels <- as.factor(df1$labels)
df1 <- df1[, !(names(df1) %in% c("ID","Over18", "EmployeeCount", "StandardHours", "Attrition"))]

train_ind <- sample(1:nrow(df1), round(.75 * nrow(df1)))
train <- df1[train_ind,]
test <- df1[-train_ind,]
```

Here I attempt to fit a KNN model including categorical variables, which results in a decent accuracy (83.49%), but very low specificity.

```
knn_model <- knncat(train, test, k = 5, classcol = 32)
confusionMatrix(as.factor(knn_model$test.classes), as.factor(test$labels))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 167  26
##          1   9  16
##
##                Accuracy : 0.8394
##                  95% CI : (0.7839, 0.8856)
##     No Information Rate : 0.8073
##     P-Value [Acc > NIR] : 0.131051
##
##                   Kappa : 0.3899
##
##  Mcnemar's Test P-Value : 0.006841
##
##             Sensitivity : 0.9489
##             Specificity : 0.3810
##          Pos Pred Value : 0.8653
##          Neg Pred Value : 0.6400
```

```
##              Prevalence : 0.8073
##          Detection Rate : 0.7661
##    Detection Prevalence : 0.8853
##       Balanced Accuracy : 0.6649
##
##        'Positive' Class : 0
##
```

```
####################knn model has very low specificity
```

```
####################attempting to fit logistic regression model
df1 <- df[complete.cases(df), ]
df1$labels <- ifelse(df1$Attrition == "No", 0, 1)
df1$labels <- as.factor(df1$labels)
df1 <- df1[, !(names(df1) %in% c("ID", "Attrition","Over18", "EmployeeCount", "StandardHours"))]
train_ind <- sample(1:nrow(df1), round(.75 * nrow(df1)))
train <- df1[train_ind,]
test <- df1[-train_ind,]
```

I now fit a logistic regression and select my explanatory variables with backward stepwise selection, utilizing AIC as a deciding metric.

```
logistic <- glm(labels ~., family = binomial(link = 'logit'), data = train) %>% stepAIC(trace = FALSE)

no_attrition <- read.csv("/Users/angelobravo/Downloads/MDS-6306-Doing-Data-Science-Fall-2019-Master-7/U
```

Here is a summary of our explanatory variable coefficients and their respective significance. Some of their p-values indicate that certain cofficient are not significant predictors not at an alpha=.05 signficance level. However, the AIC stepwise method attempts to find the best parsimonious model, with minimal bias. From this model, an accuracy of 88.53%, sensitivity of 90%, and specificity of 72.22% was attained.

```
summary(logistic)
```

```
##
## Call:
## glm(formula = labels ~ Age + BusinessTravel + DistanceFromHome +
##     EnvironmentSatisfaction + HourlyRate + JobInvolvement + JobRole +
##     JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime +
##     RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear +
##     WorkLifeBalance + YearsSinceLastPromotion + YearsWithCurrManager,
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8346  -0.4486  -0.1826  -0.0465   3.5183
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      3.409456   1.610076   2.118 0.034211 *
## Age                             -0.055143   0.021825  -2.527 0.011516 *
## BusinessTravelTravel_Frequently  2.120729   0.641389   3.306 0.000945 ***
## BusinessTravelTravel_Rarely      1.086336   0.563412   1.928 0.053838 .
## DistanceFromHome                 0.040077   0.017192   2.331 0.019747 *
## EnvironmentSatisfaction         -0.388861   0.138677  -2.804 0.005046 **
## HourlyRate                       0.011842   0.007158   1.654 0.098030 .
## JobInvolvement                  -0.888215   0.199909  -4.443 8.87e-06 ***
## JobRoleHuman Resources           2.153865   0.958162   2.248 0.024582 *
```

```
## JobRoleLaboratory Technician     1.448821   0.708615   2.045 0.040896 *
## JobRoleManager                   1.271410   0.982507   1.294 0.195649
## JobRoleManufacturing Director   -1.605784   1.218925  -1.317 0.187712
## JobRoleResearch Director        -0.570277   1.291378  -0.442 0.658776
## JobRoleResearch Scientist        0.927318   0.698345   1.328 0.184218
## JobRoleSales Executive           1.369023   0.682503   2.006 0.044868 *
## JobRoleSales Representative      2.580676   0.795086   3.246 0.001171 **
## JobSatisfaction                 -0.606661   0.135490  -4.478 7.55e-06 ***
## MaritalStatusMarried             0.719410   0.452990   1.588 0.112255
## MaritalStatusSingle              1.897399   0.468621   4.049 5.15e-05 ***
## NumCompaniesWorked               0.238805   0.060619   3.939 8.17e-05 ***
## OverTimeYes                      2.050888   0.311017   6.594 4.28e-11 ***
## RelationshipSatisfaction        -0.431704   0.132091  -3.268 0.001082 **
## TotalWorkingYears               -0.085633   0.040685  -2.105 0.035307 *
## TrainingTimesLastYear           -0.256561   0.121160  -2.118 0.034214 *
## WorkLifeBalance                 -0.633962   0.199157  -3.183 0.001456 **
## YearsSinceLastPromotion          0.286524   0.065432   4.379 1.19e-05 ***
## YearsWithCurrManager            -0.176843   0.065303  -2.708 0.006768 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 582.14  on 651  degrees of freedom
## Residual deviance: 339.95  on 625  degrees of freedom
## AIC: 393.95
##
## Number of Fisher Scoring iterations: 7
```

```
no_attrition <- no_attrition[, !(names(no_attrition) %in% c("ID","Attrition","Over18", "EmployeeCount",

results <- ifelse(predict(logistic, test, type = "response") < .5, 0, 1)
attr(results, "names") <- NULL

confusionMatrix(as.factor(test$labels), as.factor(results))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 180   5
##          1  20  13
##
##                Accuracy : 0.8853
##                  95% CI : (0.8354, 0.9244)
##     No Information Rate : 0.9174
##     P-Value [Acc > NIR] : 0.96229
##
##                   Kappa : 0.4512
##
##  Mcnemar's Test P-Value : 0.00511
##
##             Sensitivity : 0.9000
##             Specificity : 0.7222
##          Pos Pred Value : 0.9730
```

```
##          Neg Pred Value : 0.3939
##              Prevalence : 0.9174
##          Detection Rate : 0.8257
##    Detection Prevalence : 0.8486
##       Balanced Accuracy : 0.8111
##
##        'Positive' Class : 0
##
```

```
no_att_results <- ifelse(predict(logistic, no_attrition, type = "response") <.5, 0, 1)
Attrition <- ifelse(no_att_results == 0, "No", "Yes")
no_att_df <- as.data.frame(Attrition)
write.csv(no_att_df, "/Users/angelobravo/Downloads/case2PredictionsBRAVO Attrition.csv", row.names = TRU
##################################################
```

```
################REGRESSION
df2 <- df[complete.cases(df), ]
df2 <- df2[, !(names(df2) %in% c("Over18", "EmployeeCount", "StandardHours", "ID"))]

train_ind <- sample(1:nrow(df2), round(.75 * nrow(df2)))
train <- df2[train_ind,]
test <- df2[-train_ind,]
```

Here I fit a linear model to accurately assess monthly income based on 31 explanatory variables. I select my explanatory variables with forward stepwise selection, utilizing AIC as a deciding metric. Some of the p-values of the explanatory indicate that certain cofficient are not significant predictors not at a .05 signficance. However, the AIC stepwise method attempts to find the best parsimonious model, with minimal bias.On a test set, an RMSE of 1120.312 was attained.

```
linear_model <- lm(MonthlyIncome ~., data = train)

summary(linear_model)
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3787.2  -651.5    33.2   560.6  3725.6
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.535e+02  9.052e+02   0.280  0.77951
## Age                               1.682e+00  6.537e+00   0.257  0.79698
## AttritionYes                      6.042e+01  1.387e+02   0.436  0.66325
## BusinessTravelTravel_Frequently   2.602e+02  1.591e+02   1.635  0.10254
## BusinessTravelTravel_Rarely       3.977e+02  1.354e+02   2.936  0.00345
## DailyRate                         1.304e-01  1.056e-01   1.234  0.21759
## DepartmentResearch & Development  3.624e+02  5.135e+02   0.706  0.48068
## DepartmentSales                  -2.492e+02  5.333e+02  -0.467  0.64049
```

```
## DistanceFromHome                -8.666e+00  5.189e+00  -1.670  0.09540
## Education                       -4.575e+01  4.309e+01  -1.062  0.28877
## EducationFieldLife Sciences      9.233e+01  4.231e+02   0.218  0.82730
## EducationFieldMarketing          5.126e+01  4.502e+02   0.114  0.90938
## EducationFieldMedical            7.858e+00  4.241e+02   0.019  0.98522
## EducationFieldOther              1.557e+02  4.533e+02   0.344  0.73131
## EducationFieldTechnical Degree   2.542e+01  4.379e+02   0.058  0.95374
## EmployeeNumber                   6.374e-02  6.934e-02   0.919  0.35835
## EnvironmentSatisfaction         -5.714e+00  3.922e+01  -0.146  0.88423
## GenderMale                       1.642e+02  8.619e+01   1.905  0.05724
## HourlyRate                      -1.289e+00  2.134e+00  -0.604  0.54593
## JobInvolvement                   1.242e+01  6.072e+01   0.205  0.83795
## JobLevel                         2.783e+03  9.638e+01  28.871  < 2e-16
## JobRoleHuman Resources           1.685e+02  5.778e+02   0.292  0.77065
## JobRoleLaboratory Technician    -5.020e+02  1.987e+02  -2.526  0.01179
## JobRoleManager                   4.477e+03  3.115e+02  14.372  < 2e-16
## JobRoleManufacturing Director    3.026e+02  1.971e+02   1.535  0.12531
## JobRoleResearch Director         4.097e+03  2.552e+02  16.054  < 2e-16
## JobRoleResearch Scientist       -2.745e+02  1.954e+02  -1.405  0.16060
## JobRoleSales Executive           7.448e+02  3.916e+02   1.902  0.05765
## JobRoleSales Representative      1.938e+02  4.264e+02   0.454  0.64970
## JobSatisfaction                  2.395e+00  3.826e+01   0.063  0.95011
## MaritalStatusMarried             4.089e+01  1.133e+02   0.361  0.71823
## MaritalStatusSingle              2.514e+01  1.561e+02   0.161  0.87215
## MonthlyRate                     -4.340e-03  5.812e-03  -0.747  0.45544
## NumCompaniesWorked              -6.069e+00  1.914e+01  -0.317  0.75121
## OverTimeYes                      1.123e+01  9.842e+01   0.114  0.90923
## PercentSalaryHike                2.733e+01  1.812e+01   1.508  0.13196
## PerformanceRating               -4.404e+02  1.812e+02  -2.431  0.01536
## RelationshipSatisfaction        -1.941e+01  3.829e+01  -0.507  0.61236
## StockOptionLevel                 4.323e+00  6.660e+01   0.065  0.94826
## TotalWorkingYears                5.114e+01  1.220e+01   4.192 3.18e-05
## TrainingTimesLastYear            1.620e+01  3.331e+01   0.486  0.62695
## WorkLifeBalance                 -6.838e+01  5.972e+01  -1.145  0.25269
## YearsAtCompany                  -2.967e+00  1.562e+01  -0.190  0.84947
## YearsInCurrentRole               7.766e+00  2.038e+01   0.381  0.70336
## YearsSinceLastPromotion          2.659e+01  1.738e+01   1.529  0.12670
## YearsWithCurrManager            -3.294e+01  2.000e+01  -1.647  0.09999
##
## (Intercept)
## Age
## AttritionYes
## BusinessTravelTravel_Frequently
## BusinessTravelTravel_Rarely      **
## DailyRate
## DepartmentResearch & Development
## DepartmentSales
## DistanceFromHome                 .
## Education
## EducationFieldLife Sciences
## EducationFieldMarketing
## EducationFieldMedical
## EducationFieldOther
## EducationFieldTechnical Degree
```

```
## EmployeeNumber
## EnvironmentSatisfaction
## GenderMale                       .
## HourlyRate
## JobInvolvement
## JobLevel                         ***
## JobRoleHuman Resources
## JobRoleLaboratory Technician     *
## JobRoleManager                   ***
## JobRoleManufacturing Director
## JobRoleResearch Director         ***
## JobRoleResearch Scientist
## JobRoleSales Executive           .
## JobRoleSales Representative
## JobSatisfaction
## MaritalStatusMarried
## MaritalStatusSingle
## MonthlyRate
## NumCompaniesWorked
## OverTimeYes
## PercentSalaryHike
## PerformanceRating                *
## RelationshipSatisfaction
## StockOptionLevel
## TotalWorkingYears                ***
## TrainingTimesLastYear
## WorkLifeBalance
## YearsAtCompany
## YearsInCurrentRole
## YearsSinceLastPromotion
## YearsWithCurrManager             .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1041 on 606 degrees of freedom
## Multiple R-squared:  0.9536, Adjusted R-squared:  0.9501
## F-statistic: 276.7 on 45 and 606 DF,  p-value: < 2.2e-16
```

```
step.model <- stepAIC(linear_model, direction = "forward",
                      trace = TRUE)
```

```
## Start:  AIC=9104.67
## MonthlyIncome ~ Age + Attrition + BusinessTravel + DailyRate +
##     Department + DistanceFromHome + Education + EducationField +
##     EmployeeNumber + EnvironmentSatisfaction + Gender + HourlyRate +
##     JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus +
##     MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
##     PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
##     TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##     YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
##     YearsWithCurrManager
```

```
summary(step.model)
```

```
##
```

```
## Call:
## lm(formula = MonthlyIncome ~ Age + Attrition + BusinessTravel +
##     DailyRate + Department + DistanceFromHome + Education + EducationField +
##     EmployeeNumber + EnvironmentSatisfaction + Gender + HourlyRate +
##     JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus +
##     MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
##     PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
##     TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##     YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
##     YearsWithCurrManager, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3787.2  -651.5    33.2   560.6  3725.6
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.535e+02  9.052e+02   0.280  0.77951
## Age                               1.682e+00  6.537e+00   0.257  0.79698
## AttritionYes                      6.042e+01  1.387e+02   0.436  0.66325
## BusinessTravelTravel_Frequently   2.602e+02  1.591e+02   1.635  0.10254
## BusinessTravelTravel_Rarely       3.977e+02  1.354e+02   2.936  0.00345
## DailyRate                         1.304e-01  1.056e-01   1.234  0.21759
## DepartmentResearch & Development  3.624e+02  5.135e+02   0.706  0.48068
## DepartmentSales                  -2.492e+02  5.333e+02  -0.467  0.64049
## DistanceFromHome                 -8.666e+00  5.189e+00  -1.670  0.09540
## Education                        -4.575e+01  4.309e+01  -1.062  0.28877
## EducationFieldLife Sciences       9.233e+01  4.231e+02   0.218  0.82730
## EducationFieldMarketing           5.126e+01  4.502e+02   0.114  0.90938
## EducationFieldMedical             7.858e+00  4.241e+02   0.019  0.98522
## EducationFieldOther               1.557e+02  4.533e+02   0.344  0.73131
## EducationFieldTechnical Degree    2.542e+01  4.379e+02   0.058  0.95374
## EmployeeNumber                    6.374e-02  6.934e-02   0.919  0.35835
## EnvironmentSatisfaction          -5.714e+00  3.922e+01  -0.146  0.88423
## GenderMale                        1.642e+02  8.619e+01   1.905  0.05724
## HourlyRate                       -1.289e+00  2.134e+00  -0.604  0.54593
## JobInvolvement                    1.242e+01  6.072e+01   0.205  0.83795
## JobLevel                          2.783e+03  9.638e+01  28.871  < 2e-16
## JobRoleHuman Resources            1.685e+02  5.778e+02   0.292  0.77065
## JobRoleLaboratory Technician     -5.020e+02  1.987e+02  -2.526  0.01179
## JobRoleManager                    4.477e+03  3.115e+02  14.372  < 2e-16
## JobRoleManufacturing Director     3.026e+02  1.971e+02   1.535  0.12531
## JobRoleResearch Director          4.097e+03  2.552e+02  16.054  < 2e-16
## JobRoleResearch Scientist        -2.745e+02  1.954e+02  -1.405  0.16060
## JobRoleSales Executive            7.448e+02  3.916e+02   1.902  0.05765
## JobRoleSales Representative       1.938e+02  4.264e+02   0.454  0.64970
## JobSatisfaction                   2.395e+00  3.826e+01   0.063  0.95011
## MaritalStatusMarried              4.089e+01  1.133e+02   0.361  0.71823
## MaritalStatusSingle               2.514e+01  1.561e+02   0.161  0.87215
## MonthlyRate                      -4.340e-03  5.812e-03  -0.747  0.45544
## NumCompaniesWorked               -6.069e+00  1.914e+01  -0.317  0.75121
## OverTimeYes                       1.123e+01  9.842e+01   0.114  0.90923
## PercentSalaryHike                 2.733e+01  1.812e+01   1.508  0.13196
## PerformanceRating                -4.404e+02  1.812e+02  -2.431  0.01536
```

```
## RelationshipSatisfaction         -1.941e+01  3.829e+01  -0.507  0.61236
## StockOptionLevel                  4.323e+00  6.660e+01   0.065  0.94826
## TotalWorkingYears                 5.114e+01  1.220e+01   4.192 3.18e-05
## TrainingTimesLastYear             1.620e+01  3.331e+01   0.486  0.62695
## WorkLifeBalance                  -6.838e+01  5.972e+01  -1.145  0.25269
## YearsAtCompany                   -2.967e+00  1.562e+01  -0.190  0.84947
## YearsInCurrentRole                7.766e+00  2.038e+01   0.381  0.70336
## YearsSinceLastPromotion           2.659e+01  1.738e+01   1.529  0.12670
## YearsWithCurrManager             -3.294e+01  2.000e+01  -1.647  0.09999
##
## (Intercept)
## Age
## AttritionYes
## BusinessTravelTravel_Frequently
## BusinessTravelTravel_Rarely      **
## DailyRate
## DepartmentResearch & Development
## DepartmentSales
## DistanceFromHome                 .
## Education
## EducationFieldLife Sciences
## EducationFieldMarketing
## EducationFieldMedical
## EducationFieldOther
## EducationFieldTechnical Degree
## EmployeeNumber
## EnvironmentSatisfaction
## GenderMale                       .
## HourlyRate
## JobInvolvement
## JobLevel                         ***
## JobRoleHuman Resources
## JobRoleLaboratory Technician     *
## JobRoleManager                   ***
## JobRoleManufacturing Director
## JobRoleResearch Director         ***
## JobRoleResearch Scientist
## JobRoleSales Executive           .
## JobRoleSales Representative
## JobSatisfaction
## MaritalStatusMarried
## MaritalStatusSingle
## MonthlyRate
## NumCompaniesWorked
## OverTimeYes
## PercentSalaryHike
## PerformanceRating                *
## RelationshipSatisfaction
## StockOptionLevel
## TotalWorkingYears                ***
## TrainingTimesLastYear
## WorkLifeBalance
## YearsAtCompany
## YearsInCurrentRole
```

```
## YearsSinceLastPromotion
## YearsWithCurrManager              .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1041 on 606 degrees of freedom
## Multiple R-squared:  0.9536, Adjusted R-squared:  0.9501
## F-statistic: 276.7 on 45 and 606 DF,  p-value: < 2.2e-16
```

```
results <- predict(step.model, test)
attr(results, "names") <- NULL

RMSE <- sqrt(sum((results-test$MonthlyIncome)^2)/length(results))

no_sal <- read.csv("/Users/angelobravo/Downloads/MDS-6306-Doing-Data-Science-Fall-2019-Master-7/Unit 14

Results <- predict(step.model, no_sal)
attr(Results, "names") <- NULL
results_df <- as.data.frame(Results)
write.csv(results_df, "/Users/angelobravo/Downloads/case2PredictionsBRAVO Salary.csv", row.names = TRUE

paste("RMSE: ", RMSE)
```

```
## [1] "RMSE:  1120.31181070441"
```