Can my child be the next Newton?

2017

ASSISTments Data Mining Competition

https://sites.google.com/view/assistmentsdatamining/home

CHALLENGE:

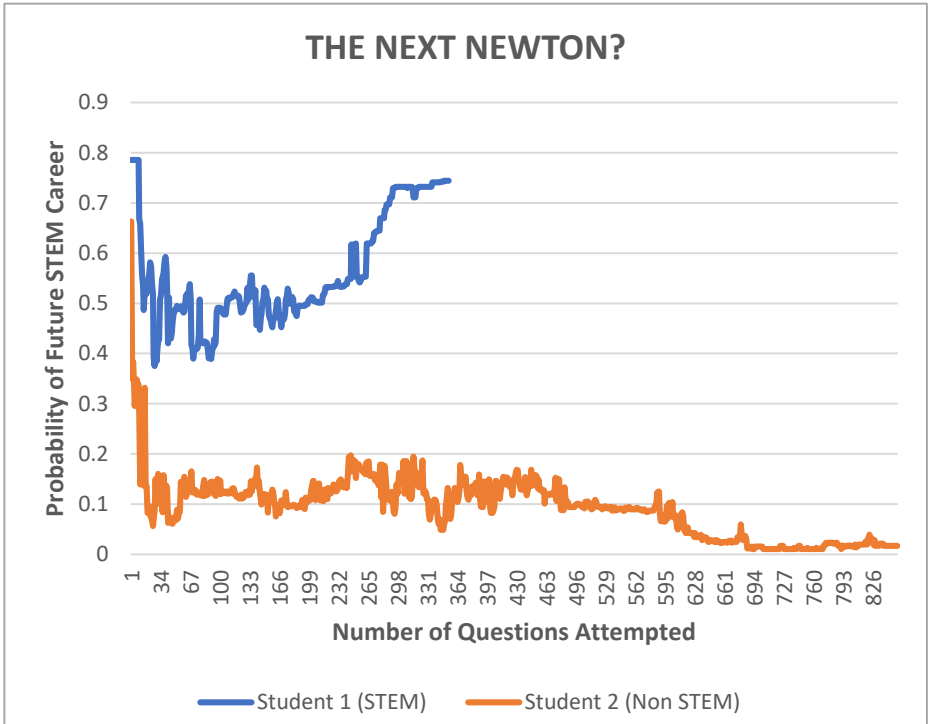
Identification of detectors for Early Warning Systems in Online Courses to predict if 'students are losing interest in STEM'

SOLUTION:

THE NEXT NEWTON™ ALGORITHM

0.
Contents

1. Preface

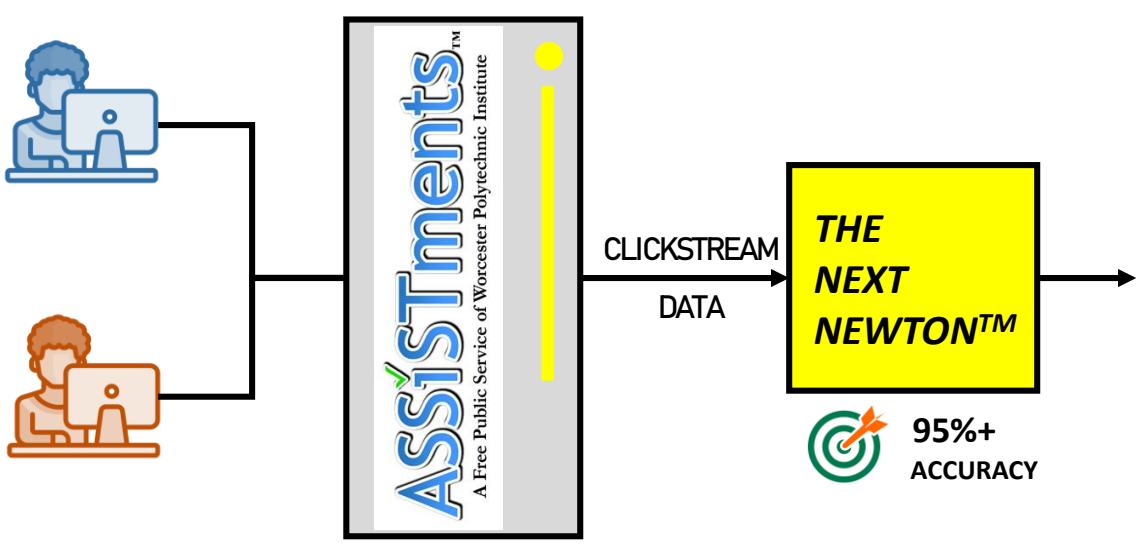


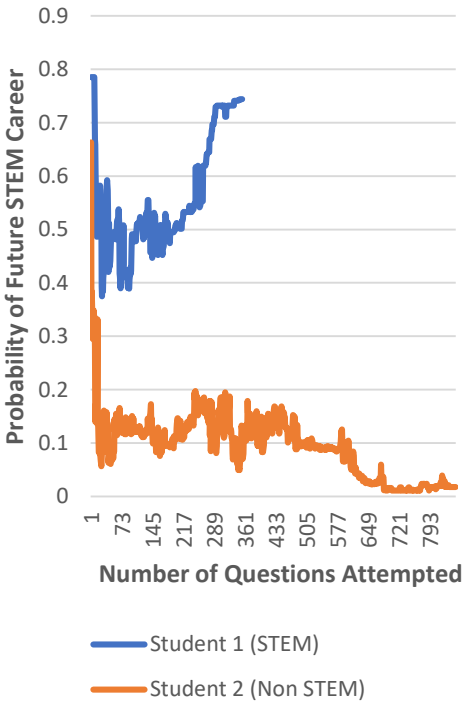
Actual Predictions by THE NEXT NEWTON™ algorithm
Real-Time probabilities of future STEM Careers
using simulated clickstream data

1. PREFACE

We can look at students' real-time performance in a specific middle high school math test and predict whether they will opt for STEM Careers after 10 years – using *THE NEXT NEWTON™* algorithm. It can predict if a child is losing interest in STEM or is not completely off the track of probably being the next Isaac Newton.

We could achieve a 95%+ prediction accuracy, which also allowed us to surpass the top score in the data-mining competition.





1. Students take math test on ASSISTments Platform

2. ASSISTments tracks student performance and engagement data

3. Our Prediction Algorithm uses past data to predict probability of future STEM career outcomes for students.

2.
Introduction

2. INTRODUCTION

What is ASSISTments?

- ASSISTments is an online tutoring platform connecting students, teachers and administrators.
- ASSISTments enables tracking of student activity online, and the 2017 Data Mining Competition used anonymized data from a specific ASSISTments case study.
- The outcome of this analysis is extremely relevant to institutions using ASSISTments for teaching. This includes over 600 teachers from 42 states and 14 countries whose students solved 10,000,000 problems last year.

http://www.aboutus.assistments.org/

What is the competition objective?

The task in this competition is to develop a cross-validated prediction model that is able to use middle-school data to predict whether the students (who have now finished college) pursue a career in STEM fields (1) or not (0).

Schools are already using 'drop out' detectors for early warning system, but they also need early warning systems for 'these students are losing interest in STEM' detectors. ' The results of this competition could help inform the design of systems that could help try to reignite student's interest in studying STEM.

This competition uses data from a longitudinal study, now over a decade long, led by Professor Ryan Baker and Professor Neil Heffernan.

2. INTRODUCTION

Assumption Alert

Before we start the analysis, our assumption is that the performance in ASSISTments has some influence over future career outcomes.

This assumption is based on the following research papers:

1. Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, 1(1), 107–128.
    - Around 3 district's data was compiled, for which affect labels were achieved using BROMP and a logistic linear regression was developed on it. The engagement of students suggested that they will be taking the STEM course and Disengagement showed they will not, however, some measures gave negative Correlation.
2. San Pedro, M., Baker, R., Bowers, A. & Heffernan, N. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In S. D'Mello, R. Calvo, & A. Olney (Eds.) Proceedings of the 6th International Conference on Educational Data Mining
    - College enrollment of students has been predicted by running multiple linear regressions, where the logistic model predicts the logit, natural logarithm of an odds ratio of an outcome variable from a predictor or set of predictors.
3. San Pedro, M., Ocumpaugh, J., Baker, R., & Heffernan, N. (2014) Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In John Stamper et al. (Eds) Proceedings of the 7th International Conference on Educational Data Mining. pp. 276-279.
    - Gaming in scaffolding questions has also been extensively studied. Gaming means randomly selecting answer choices and not focusing on the question either due to boredom or higher proficiency.

3.
Data Overview

3. DATA OVERVIEW

What types of data were available?

The dataset for this competition followed students for about 10 years, starting when they used an online mathematics system as middle-school students (from 2004-2006) through to when they graduated from college.

418 megabytes of clickstream data of students answering questions on an online testing platform was available for analysis. **Career outcomes of 514 students** are provided as isSTEM = '0' or '1'.

https://sites.google.com/view/assistmentsdatamining/winners-of-the-2017-competion

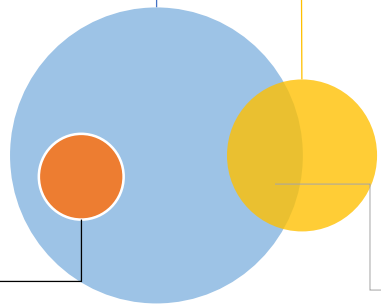
2004 – 2006
Student Generated
Clickstream Data

10 Years

Student Career Outcomes

- 942,816 Observations and 76 Numeric Variables
- 1709 Unique Students have multiple 'clicks' or responses to questions recorded

- 514 Students and 4 variables
- isSTEM is the response variable with two classes '0' or '1' : Indicating No or Yes for a current STEM career



TEST DATA
- 172 Students
- 91,038 Observations and 79 Variables
- **Prediction Task: Predict isSTEM**

LABELLED DATA (TRAIN DATA)
- 467 Students
- 272,562 observations and 80 variables

3. DATA OVERVIEW

What is the Prediction Task?

This is a classification problem.
We have to classify the 172 students as having a STEM career or not using the available data.

Understanding the Data

Both Training and Test data have multiple records (500+ on an average) for a single student.
Also, every record has 76 fields. Understanding the following key ideas helped us get around the dataset:

1. Every record is a single attempt on a question by a student. So if there are 500 records for a student, every record has data for one attempt.
2. The 76 fields mainly describe the student emotional states, the question type and the description of response of the student on the question.
    - **For example:** Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.* Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
3. Each attempt could be a correct or an incorrect response to a question.
4. Each question is a part of an ASSISTment – which is problem that can be broken down into smaller problems called Scaffoldings – and a number of ASSISTments make up an assignment.
5. A student can also take hints, and there is a finite number of hints in a question before a student reaches the bottom hint.
6. A student can also take hints, and there is a finite number of hints in a question before a student reaches the bottom hint.

What next? (Challenges)

As every student has on an average 500 records, there are almost (500*76) **3800** dimensions available for every student.
This would be very difficult for a classification model to handle, so **we will have to create usable features** for the models.

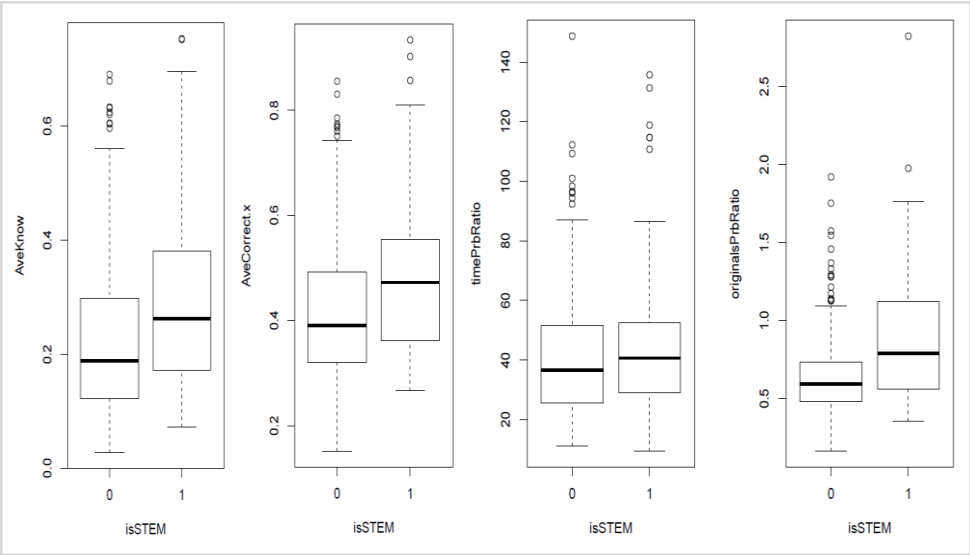
4.
Identifying Features

4. IDENTIFYING FEATURES

How did we create features?

The dataset was required to be condensed into a format wherein every observation in the dataset was of a unique student id. We created 29 unique features that showed clear trends for different classes.

- We created pivot tables which had features for every unique student using data.table()
- Logical aggregates like MAX, MIN, MEAN for different columns were created for columns that made sense according to the data dictionary.
- We created boxplots to see the distribution of the feature values between students with STEM and non-STEM careers.
- Correlation heatmaps allowed us to remove redundant features with high correlation.



Boxplot Analysis examples showing differences in different classes

	AveKnow	AveCarelessness	AveCorrect.x	totalAttempts	AveResBored	AveResEngcon	AveResConf	AveResFrust	AveResOfftask	AveResGaming	Time	Correct
AveKnow	1.00	0.96	0.92	-0.09	-0.08	0.21	-0.31	-0.19	-0.01	-0.43	0.36	0.36
AveCarelessness	0.96	1.00	0.82	-0.02	-0.14	0.22	-0.25	-0.08	-0.09	-0.39	0.38	0.41
AveCorrect.x	0.92	0.82	1.00	-0.27	0.09	0.13	-0.27	-0.23	0.14	-0.57	0.24	0.18
totalAttempts	-0.09	-0.02	-0.27	1.00	-0.73	0.50	-0.28	0.15	-0.69	0.66	0.51	0.75
AveResBored	-0.08	-0.14	0.09	-0.73	1.00	-0.72	0.48	-0.10	0.87	-0.70	-0.29	-0.60
AveResEngcon	0.21	0.22	0.13	0.50	-0.72	1.00	-0.34	-0.13	-0.63	0.39	0.22	0.48
AveResConf	-0.31	-0.25	-0.27	-0.28	0.48	-0.34	1.00	0.02	0.35	-0.35	-0.18	-0.29
AveResFrust	-0.19	-0.08	-0.23	0.15	-0.10	-0.13	0.02	1.00	-0.17	0.07	0.06	0.08
AveResOfftask	-0.01	-0.09	0.14	-0.69	0.87	-0.63	0.35	-0.17	1.00	-0.54	-0.34	-0.56
AveResGaming	-0.43	-0.39	-0.57	0.66	-0.70	0.39	-0.35	0.07	-0.54	1.00	0.05	0.27
Time	0.36	0.38	0.24	0.51	-0.29	0.22	-0.18	0.06	-0.34	0.05	1.00	0.85
Correct	0.36	0.41	0.18	0.75	-0.60	0.48	-0.29	0.08	-0.56	0.27	0.85	1.00
original	0.37	0.43	0.19	0.68	-0.58	0.43	-0.28	0.17	-0.52	0.25	0.78	0.94
Hint	-0.32	-0.28	-0.47	0.88	-0.66	0.42	-0.32	0.06	-0.58	0.82	0.36	0.58
Scaffold	0.00	0.07	-0.14	0.84	-0.58	0.46	-0.12	0.07	-0.61	0.40	0.75	0.86
Bottomhint	-0.34	-0.31	-0.48	0.83	-0.66	0.41	-0.31	0.04	-0.55	0.87	0.29	0.51
frhelpPrbRatio	0.05	0.10	-0.10	0.87	-0.65	0.49	-0.27	0.06	-0.63	0.49	0.75	0.89

Heatmap Analysis helped remove correlated features and avoid overfitting

4. IDENTIFYING FEATURES

Selecting Relevant Features

Criterion similar to the following were used to identify **29 important features**.

- The students who pursued STEM career, in general, attempted more original questions, had the higher average knowledge, a higher average of correct answers, attempted higher number of problem types and a higher number of attempts compared to the students who did not pursue STEM.
- The students who did not pursue STEM career, in general, had higher frustration, higher gaming, higher hints taken per problem compared to the students who pursued STEM
- Certain parameters had more or less same median for both STEM and Non-STEM fields, but it had different interquartile range. For example, scaffolding per problem ratio had the shorter interquartile range for Non-STEM students than for STEM students, indicating that STEM students have wider distribution in scaffolding.

```
(isSTEM = max(isSTEM),
 AveKnow = mean(AveKnow),
 AveCarelessness = mean(AveCarelessness),
 totalAttempts = mean(NumActions),
 AveResBored = mean(AveResBored),
 AveResEngcon = mean(AveResEngcon),
 AveResConf = mean(AveResConf),
 AveResFrust = mean(AveResFrust),
 AveResOfftask = mean(AveResOfftask),
 AveResGaming =  mean(AveResGaming),
 Time = sum(timeTaken),
 Correct = sum(correct),
 original = sum(original),
 Hint = sum(hint),
 Scaffold = sum(scaffold),
 Bottomhint = sum(bottomHint),
 frhelpPrbRatio = sum(frIsHelpRequest),
 stlHintUsed = sum(stlHintUsed),
 WorkingatSchool = sum(frWorkingInSchool),
 Responsefilling = sum(responseIsFillIn),
 endsWithScaffolding = sum(endsWithScaffolding),
 endsAutoScaffolding = sum(endsWithAutoScaffolding),
 frTimeTakenOnScaffolding = mean(frTimeTakenOnScaffolding),
 frIsHelpRequestScaffolding = sum(frIsHelpRequestScaffolding),
 timeGreater5Secprev2wrong = sum(timeGreater5Secprev2wrong),
 helpAccessUnder2 = sum(helpAccessUnder2Sec),
 timeGreater10SecAndNextActionRight = sum(timeGreater10SecAndNe
 timeOver80 = sum(timeOver80),
 manywrong = sum(manywrong)
```

29 Relevant features created in R based on initial analysis

5.
Modeling Approach

5. MODELING APPROACH

What models does the *THE NEXT NEWTON™* use ?

THE NEXT NEWTON™ algorithm uses a random forest algorithm. It achieves a 95.3% accuracy through just five important features selected using elasticnet penalty in logistic regression with the previously created features.

Without using the selected features, the random forest algorithm gave us an accuracy of 93% with 30 features.

```
> rownames(tabRFtest) <- c("Predicted Non-STEM", "Predicted STEM")
> colnames(tabRFtest) <- c("Actual Non-STEM", "Actual STEM")
> RFaccuracy <- sum(diag(tabRFtest))/sum(tabRFtest)
> RFaccuracy
[1] 0.9534884
> tabRFtest
                  Actual Non-STEM Actual STEM
Predicted Non-STEM             116           8
Predicted STEM                   0          48
>
```

THE NEXT NEWTON™
Algorithm Accuracy

Important Features selected using Elasticnet model

5. MODELING APPROACH

What is the Random Forest Algorithm?

- It is a supervised classification algorithm. As the name suggests, it creates the forest by some way and randomly selects the features.
- Random Forest algorithm works on the concept of bootstrapping, wherein it creates multiple number of trees, and every tree is different from one another which is because of its randomly selected features. The larger the number of trees, the more accurate our result is.

ADVANTAGES

- Random Forest algorithm avoids the issue of overfitting in the classification problems
- It can be used in both the classification and regression problems
- It can be used in filtering out the important features and thus, helps in feature engineering.

LIMITATIONS

Multiple decision trees are used in the model, which can lead to losing the interpretability of the final model.

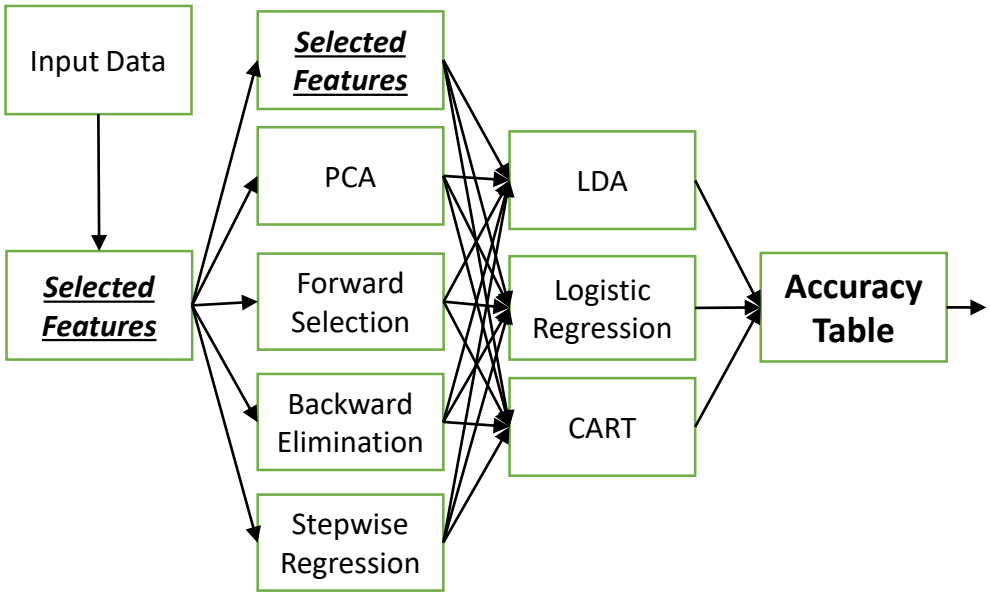
OVERCOMING LIMITATIONS: REACHING THE FINAL FEATURES

We fed 30 features to the algorithm, and then gradually started removing the features that did not reduce the accuracy. Eventually, we were able to increase the accuracy while reducing the features to just 5. We experimented using the features highlighted by the ridge, lasso and elasticnet models – and achieved the highest accuracy by elasticnet features.

5. MODELING APPROACH

Did we use any other algorithms?

Yes, WE TRIED OUT 120+ COMBINATIONS. (*We have not mentioned the models and features that did not give favourable results*)
We implemented multiple models with various subsets in it. This enabled us to sift through the important features and feed it in Random Forest which gave us the best accuracy.



Sr No.	Input Data	Model	Split Ratio	Accuracy
1	Selected Features	LDA	0.6	0.8152
2	Selected Features	Logistic	0.6	0.7935
3	Selected Features	CART models	0.6	0.7283
4	Selected Features	Ridge	0.6	0.8043
5	Selected Features	Elasticnet	0.6	0.7989
6	PCA	LDA	0.6	0.8043
7	PCA	Logistic	0.6	0.7880
8	PCA	CART models	0.6	0.7228
9	PCA	Ridge	0.6	0.8043
10	PCA	Elasticnet	0.6	0.7989
11	Stepwise Regression	LDA	0.6	0.8152
12	Stepwise Regression	Logistic	0.6	0.8152
13	Stepwise Regression	CART models	0.6	0.7554
14	Stepwise Regression	Ridge	0.6	0.8043
15	Stepwise Regression	Elasticnet	0.6	0.7989
16	Forward Selection	LDA	0.6	0.8152
17	Forward Selection	Logistic	0.6	0.7935
18	Forward Selection	CART models	0.6	0.7283
19	Forward Selection	Ridge	0.6	0.8043
20	Forward Selection	Elasticnet	0.6	0.7989
21	Backward Elimination	LDA	0.6	0.8152
22	Backward Elimination	Logistic	0.6	0.8152
23	Backward Elimination	CART models	0.6	0.7554
24	Backward Elimination	Ridge	0.6	0.8043
25	Backward Elimination	Elasticnet	0.6	0.7989

Feature Selection Process | Model Selection Process

6.
Analysis Results

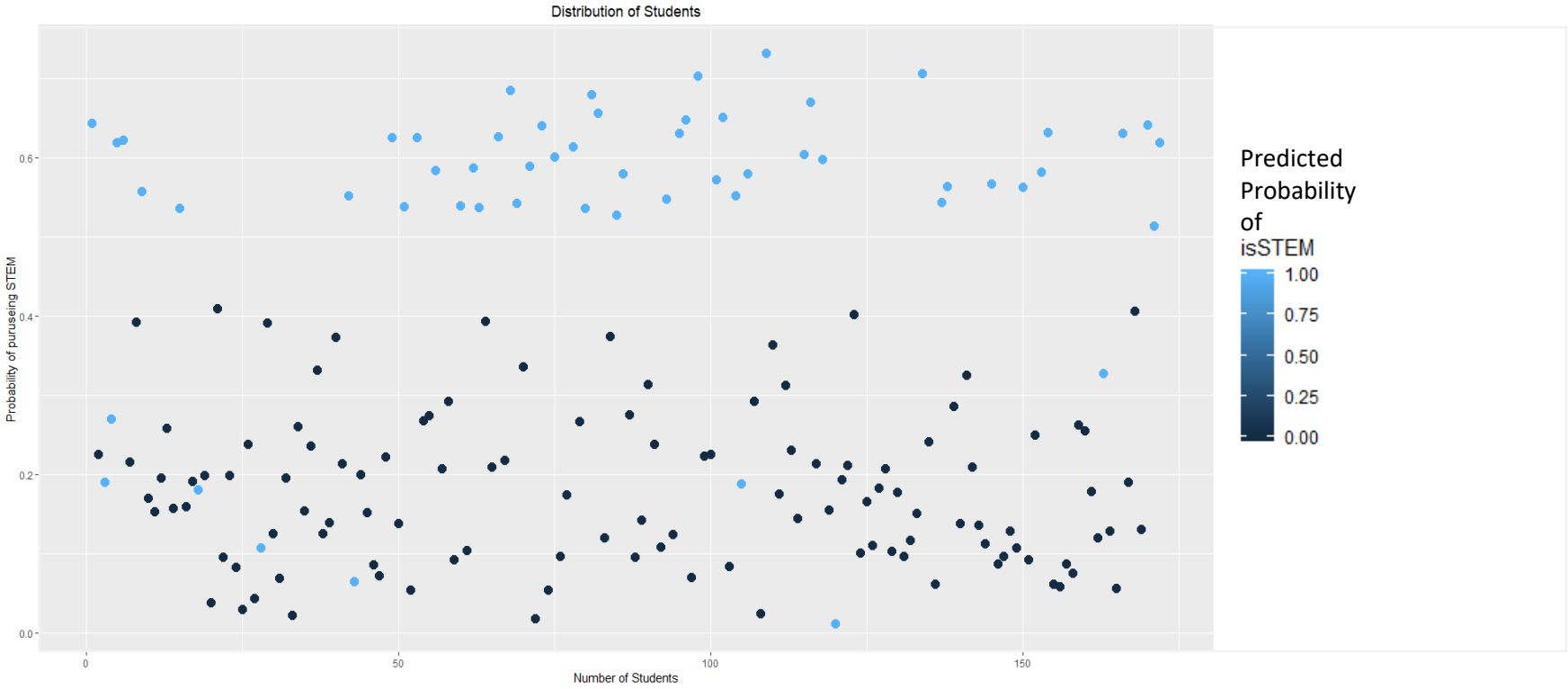
6. ANALYSIS RESULTS

Is your model a good classifier?

The model is very effective in classifying the students according to isSTEM variable. We can see that only 8 students were misclassified out of 172 students resulting an accuracy of 95.34%.

```
> RFaccuracy
[1] 0.9534884
> tabRFtest

                    Actual Non-STEM Actual STEM
  Predicted Non-STEM             116           8
  Predicted STEM                   0          48
>
```

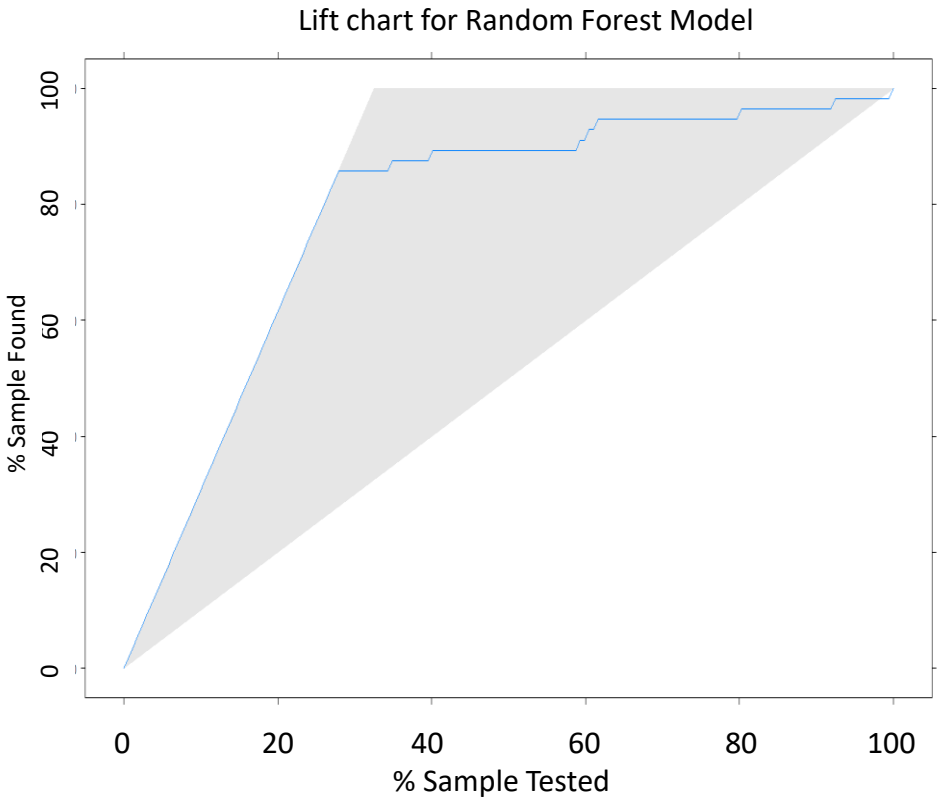


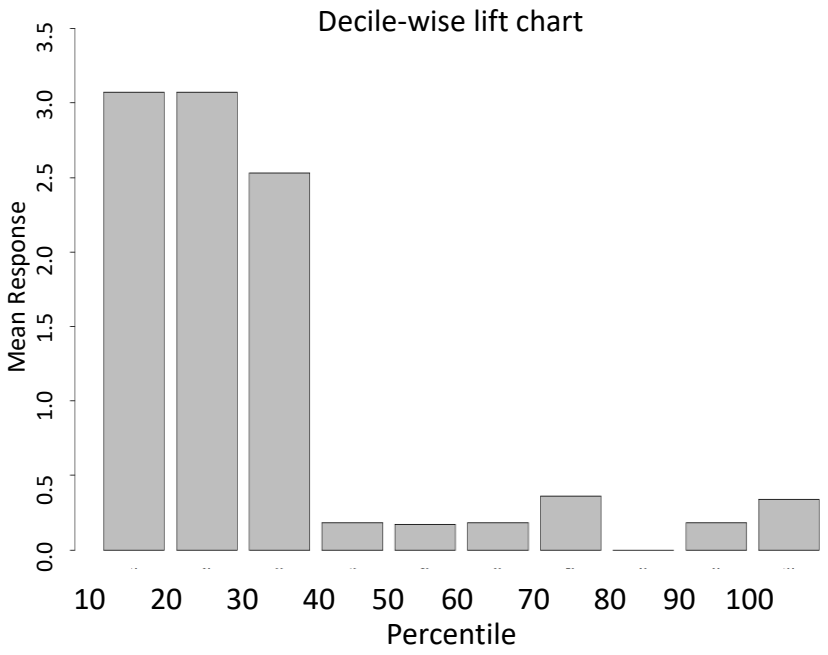
6. ANALYSIS RESULTS

How effective is the model?

The lift chart tells us that by choosing the top 30% of the data with the highest propensities, we can reach 85.71% of the STEM students. It generates a high lift of 1.85, which means that our Random Forest model is 1.85 times better than the Naïve Bayes model.

Also, in the Decile-wise lift chart, we obtained high mean response rate of 3 for the 20 percentile and 30th percentile has a mean response of 2.5 which suggests that this model outperforms Naïve Bayes.

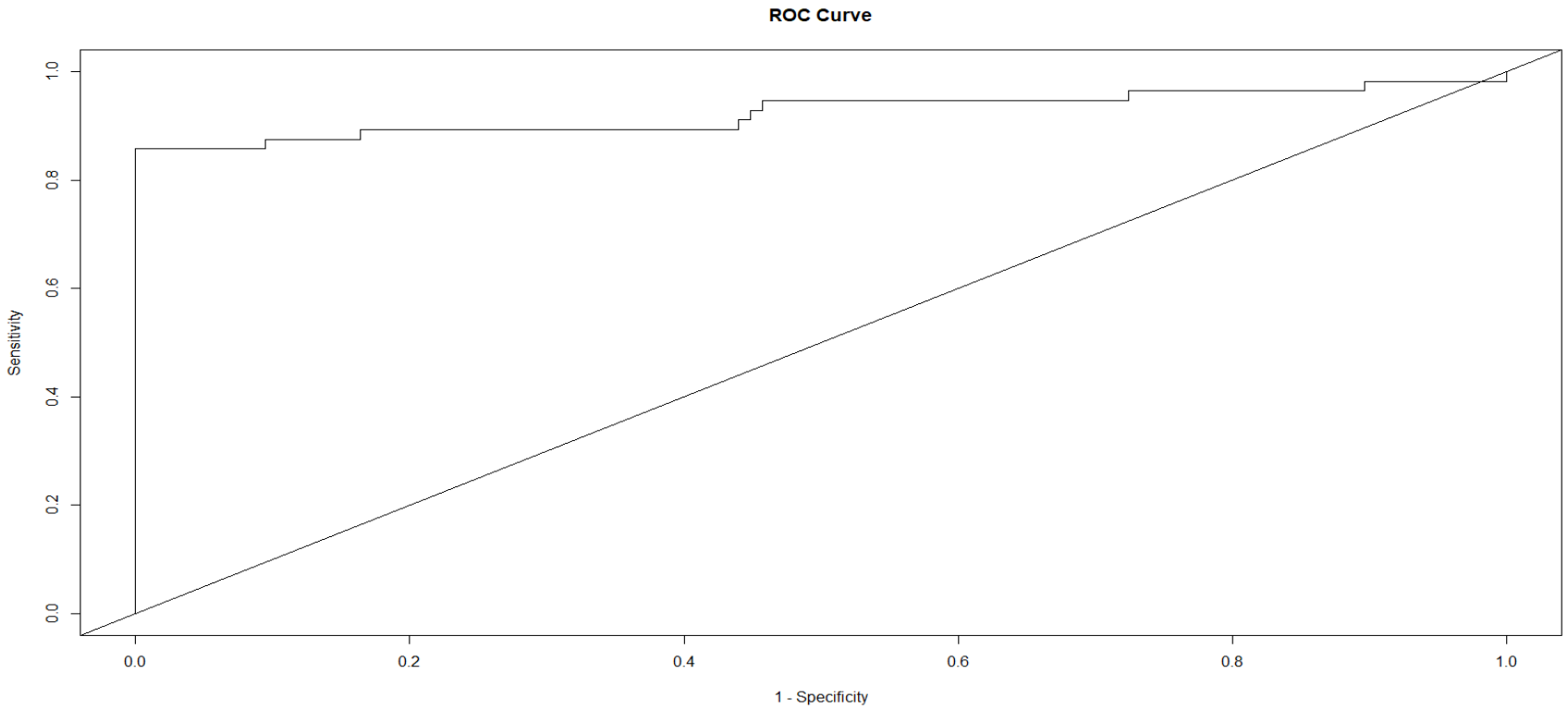




6. ANALYSIS RESULTS

How effective is the model?

The AUC is calculated to be 0.925, which indicates that the Random Forest model represents a good measure of separability. It also means that there is a 92.5% chance that the model will be able to distinguish between the positive class and the negative class.



7. Recommendations

7. RECOMMENDATIONS

What do the results mean to the educators/researchers using ASSISTments?

We have two suggestions that can be of use to people designing/administering ASSISTments tests.

1. Driving Research

The predictive model identifies five unique features that explain 95% of the students' career trajectories.

Learning what factors influence these five variables could help in design of tests that help to motivate more students to take up STEM.

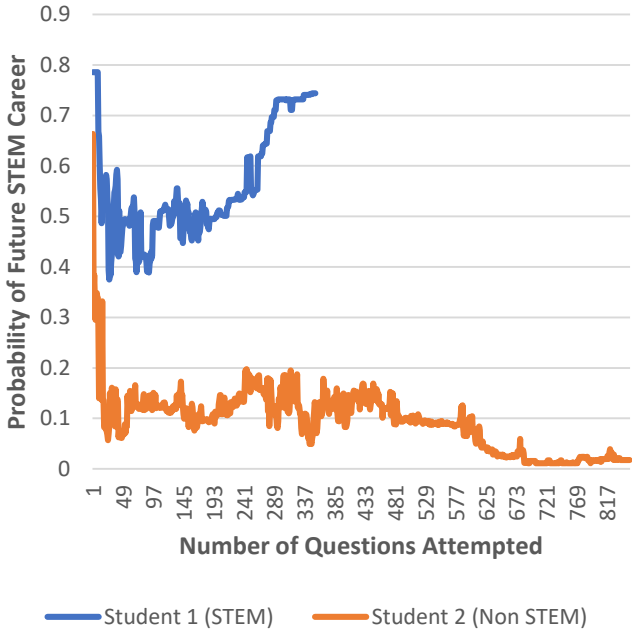
Feature
AveCarelessness
AveResFrust
Mean(hint)
Mean(original)
sum(frIsHelpRequest)/max(totalFrAttempted)

2. Real Time Performance Prediction

The five factors that help identify probability of stem career are calculated values of student performance over a period of time.

The probability of a student taking up STEM can be calculated with every response of the student, and an analysis of the timeline of two students shows a clear trend in the graph on the right.

Such a metric can help teachers identify disinterested students and possibly try to assess the reasons for the same.



8.
References

8. REFERENCES

- Paper reporting on affect measures available in the data set:
  San Pedro, M., Baker, R., Gowda, S., & Heffernan, N. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In Lane, Yacef, Mostow & Pavlik (Eds) *The Artificial Intelligence in Education Conference.* Springer-Verlag. pp. 41-50.
- First prediction paper, showing that ASSISTments data (and the variables available in this data set) predict state test scores:
  Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, 1(1), 107–128.
  First appeared as Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2013) Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, 117-124.
- Second prediction paper, showing that these variables predict who enrolls in college several years after using ASSISTments:
  San Pedro, M., Baker, R., Bowers, A. & Heffernan, N. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In S. D'Mello, R. Calvo, & A. Olney (Eds.) Proceedings of the 6th International Conference on Educational Data Mining.
- Third prediction paper, showing that these variables can predict college major:
  San Pedro, M., Ocumpaugh, J., Baker, R., & Heffernan, N. (2014) Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In John Stamper et al. (Eds) Proceedings of the 7th International Conference on Educational Data Mining. pp. 276-279.
- Additional paper, exploring role of gaming the system in college major:
  San Pedro, M.O., Baker, R., Heffernan, N., Ocumpaugh, J. (2015) Exploring College Major Choice and Middle School Student Behavior, Affect and Learning: What Happens to Students Who Game the System? Proceedings of the 5th International Learning Analytics and Knowledge Conference. pp 36-40.
- Paper exploring the degree to which affect models generalize across students from urban, suburban, and rural areas:
  Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014) Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology, 45* (3), 487-501. OI: 10.1111/bjet.12156

8. REFERENCES

- Recent papers presenting enhancements to affective models:

  Wang, Y., Heffernan, N, & Heffernan, C. (2015) Towards better affect detectors: effect of missing skills, class features and common wrong answers. Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. pp 31-35. See data here and here.

  Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017, June). Improving Sensor-Free Affect Detection Using Deep Learning. *Proceedings of the Eighteenth International Conference on Artificial Intelligence in Education* .

- If you are interested in the method used to collect measures of behavior and affect (which were used to create affect models), you may want to look at the BROMP training manual.

  Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.* Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

THE SECRET SAUCE
(R CODE)

```
### read the data

train_all <- read.csv("labelled_training_data.csv")
test_all <- read.csv("labelled_test_data.csv")

## function to create features

features_1 <- function(df){
  library(data.table)
  dft <- as.data.table(df)
  df_piv1 <- as.data.frame(dft[,.(isSTEM = max(isSTEM),
                                  AveCarelessness =
mean(AveCarelessness),
                                  AveResFrust = mean(AveResFrust),
                                  hint = mean(hint),
                                  original = mean(original),
                                  frIsHelpRequest =
sum(frIsHelpRequest)/max(totalFrAttempted)
  ),by=list(ITEST_id)])

  return(df_piv1)
}
```

THE SECRET SAUCE
(R CODE)

```
## Creating features

df_piv1 <- features_1(train_all)
test_piv1 <- features_1(test_all)

# Storing the result in test_result and removing the isSTEM variable
from test_piv1
test_result <- test_piv1[,c(1,2)]
test_piv1[,2] <- NULL

## running the algorithm

library(randomForest)
Rf_train <- randomForest(formula = isSTEM ~., data = df_piv1[,-1],
importance = TRUE)
Rf_test <- predict(Rf_train, test_piv1[,-(1)], type = "class")
```

THE SECRET SAUCE
(R CODE)

```
# Confusion Matrix for Random Forest

tabRFtest <- table(Rf_test>0.5, test_result$isSTEM)
rownames(tabRFtest) <- c("Predicted Non-STEM", "Predicted STEM")
colnames(tabRFtest) <- c("Actual Non-STEM", "Actual STEM")
tabRFtest
RFaccuracy <- sum(diag(tabRFtest))/sum(tabRFtest)
RFaccuracy

RF.df <- as.data.frame(Rf_test)
RF.df[,2] <- test_result[2]

# Distribution of Students
RF.df$Order <- 1:nrow(RF.df)

library(ggplot2)

ggplot(RF.df, aes(x=Order, y=Rf_test, col=isSTEM)) + geom_point(size = 3.5) + labs(x ="Number
of Students",
        y = "Probability of puruseing STEM", title = "Distribution of Students") +
        theme(plot.title = element_text(hjust = 0.5))
```

THE SECRET SAUCE
(R CODE)

```
# Lift Chart
RF.df <- RF.df[order(RF.df$Rf_test, decreasing = T),]

library(caret)
lift.test <- lift(relevel(as.factor(isSTEM),ref="1")~Rf_test,data=RF.df)
xyplot(lift.test,plot="gain", main = "Lift chart for the Random Forest model")

# Decile Chart
RF.df$isSTEM <- as.numeric(RF.df$isSTEM)

library(gains)
gain <- gains(RF.df$isSTEM, RF.df$Rf_test)
barplot(gain$mean.resp / mean(RF.df$isSTEM), names.arg = gain$depth, xlab = "Percentile",
      ylab = "Mean Response", main = "Decile-wise lift chart", ylim = c(0,3.5))
```

THE SECRET SAUCE
(R CODE)

```
# Accuracy vs Cutoff graph
library(ROCR)
pred <- prediction(Rf_test, test_result$isSTEM)
eval <- performance(pred,"acc")
plot(eval, main = "Accuracy vs Cutoff graph")
abline(v=0.5)

# ROC
roc <- performance(pred,"tpr", "fpr")
plot(roc,
    main="ROC Curve",
    ylab = "Sensitivity",
    xlab = "1 - Specificity")
abline(a=0,b=1)

# Calculating AUC
auc <- performance(pred,"auc")
auc <- unlist(slot(auc,"y.values"))
auc
```