Abstract

Background: The client provided survey & logistic data including a binary recommend/don't recommend and background information.

Our study aimed to find the best measures of program success beyond the fact whether students recommended the course or not

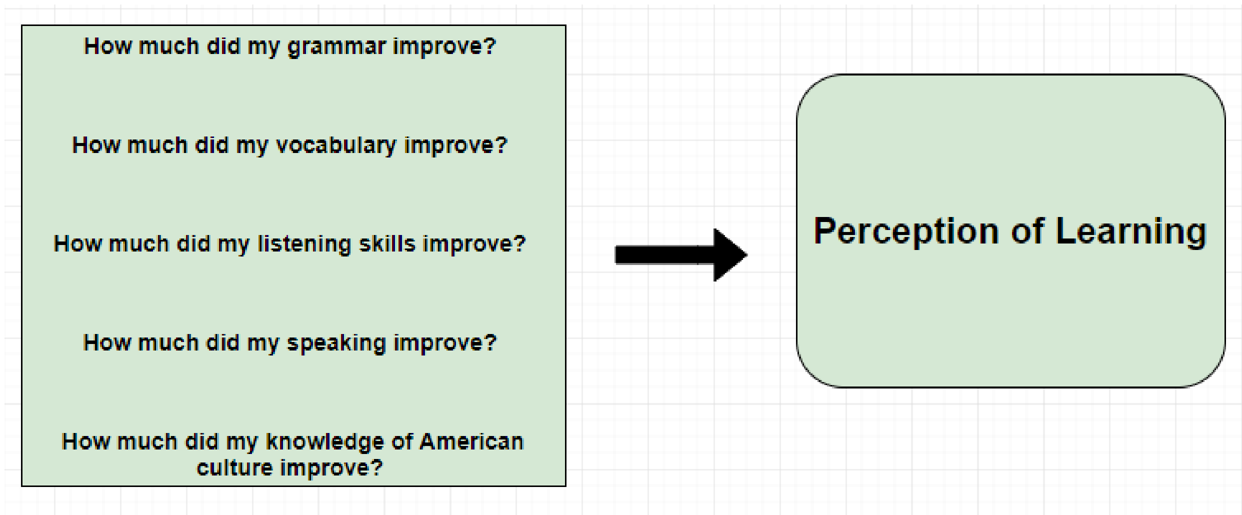
Survey questions were analyzed and grouped into a new metric variable called "Perception of Learning"; then statistical analyses were performed measuring the impact of variables such as "Age", "Gender", "Living Situation" on "Perception of Learning"

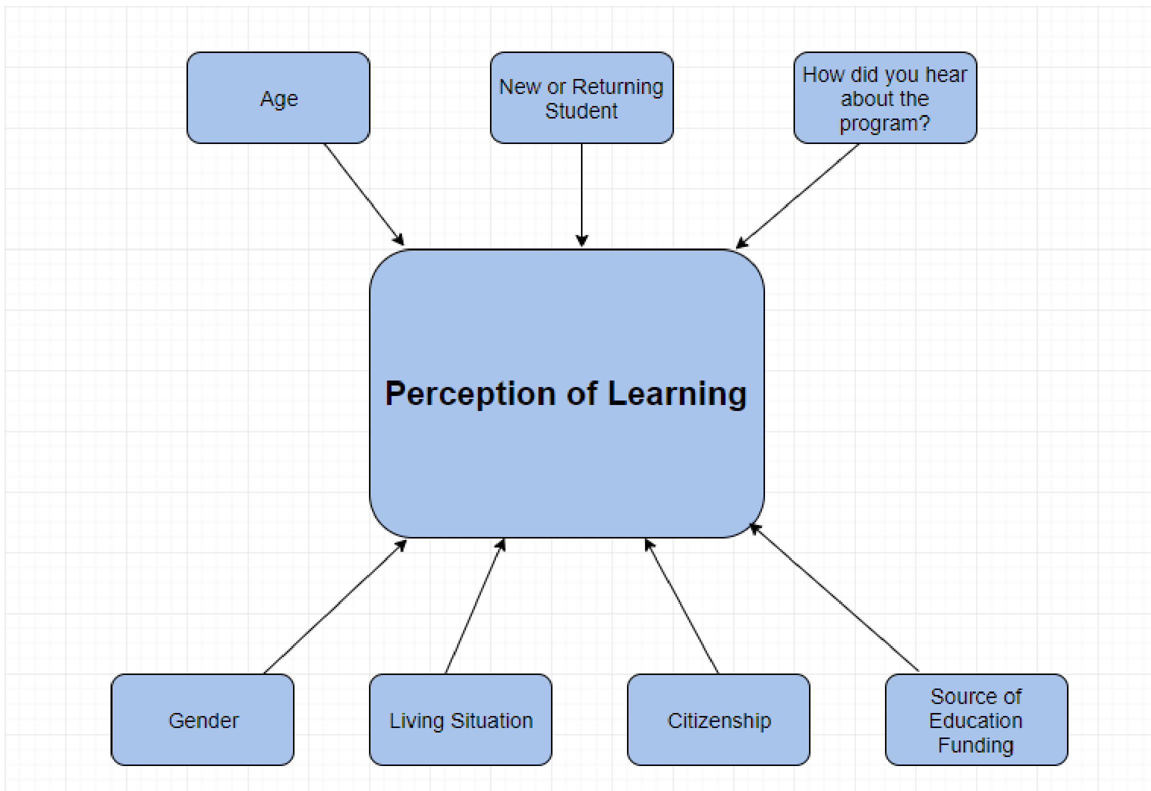
Our final model consists of three variables: 'Age', 'Living Conditions', and 'Source of funding'; these variables were most important in predicting the student's perception of learning from the ALC program. The model had a 61.6% accuracy using these three variables and was able to correctly identify 24 out of the 29 people who wouldn't recommend the ALC program.

Description of Variables

ID: Student ID -categorical
Date - start date of program - categorical
Quarter - quarter of program - categorical
Recommend: Do you recommend this program? (yes/no)- categorical
Grammer: How much did my grammar improve? (A, B, C, D) - ordinal
Vocab: How much did my vocabulary improve? (A, B, C, D) - ordinal
Listening: How much did my listening skills improve? (A, B, C, D) - ordinal
Speaking: How much did my speaking improve? (A, B, C, D) - ordinal
Culture: How much did my knowledge of American culture improve? (A, B, C, D) - ordinal

We want the responses to these questions analyzed as a function of the relevant predictors.
Potential predictors could be:
Age - numerical
Citizenship: Citizenship_U - categorical
Gender - categorical
Major - categorical
Source_of_Funding: Who pays for you to study at ALC - categorical
How_did_you_hear: How did you hear about ALC - categorical
Teacher_rate - ordinal
Living: In_my_opinion_my_housing_was - ordinal
Status: New or returning(status) - categorical





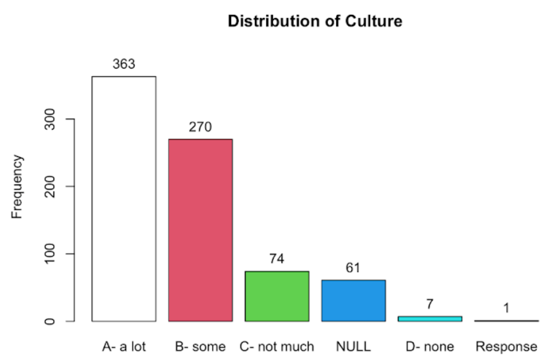
Description of Data

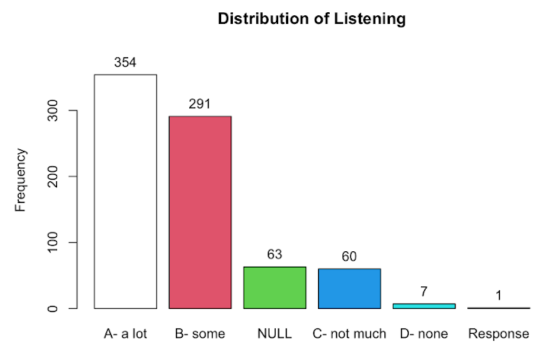
The client provided survey and logistical data from the UCLA Extension program.
We used various program evaluation datasets for self-evaluation of improvement, and we utilized demographic and background data of the students. The dataset was interesting because there were many survey questions that promoted open-ended responses. Initially, we had about 80 variables just from program evaluation surveys, but narrowed focus to mostly grammar, speaking, vocabulary, American culture, and demographic information. Our methodology began after we combined 5 datasets and removed invalid or null values from the predictors. We began with 776 observations, but with cleaning, our final dataset consisted of 707 observations. This

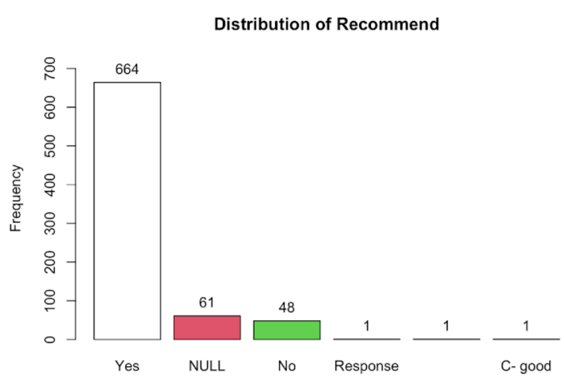
established the data that we would use, manipulate and create our own metrics for studying program success.

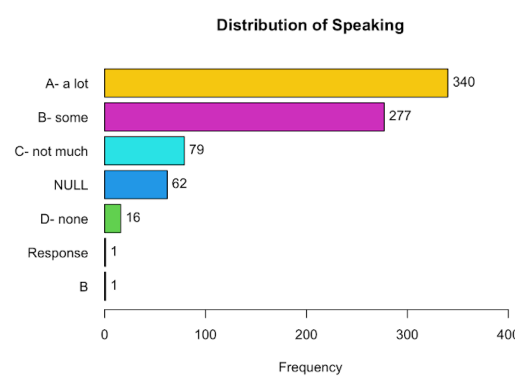
Research Question
What are our best measures of program success, beyond the binary recommend/don't recommend response? To answer this we had to define program success and study factors are relevant to program success.

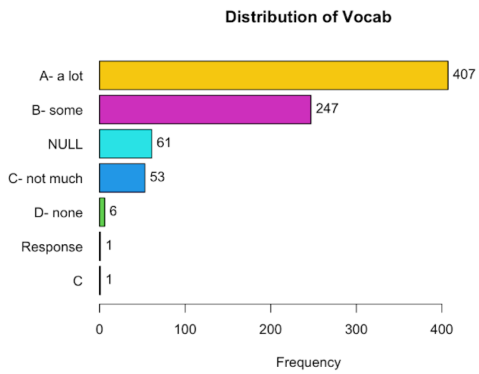
Exploratory Data Analysis

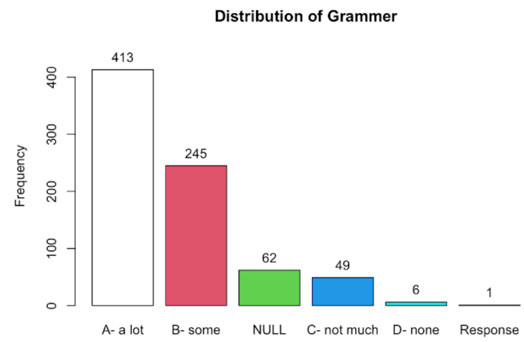












Speaking Frequency Table

	Frequency	Percent	Cum. percent
A- a lot	340	43.8	43.8
B- some	277	35.7	79.5
C- not much	79	10.2	89.7
NULL	62	8.0	97.7
D- none	16	2.1	99.7
Response	1	0.1	99.9
B	1	0.1	100.0
Total	776	100.0	100.0

Grammer Frequency Table

	Frequency	Percent	Cum. percent
A- a lot	413	53.2	53.2
B- some	245	31.6	84.8
NULL	62	8.0	92.8
C- not much	49	6.3	99.1
D- none	6	0.8	99.9
Response	1	0.1	100.0
Total	776	100.0	100.0

Vocab Frequency Table

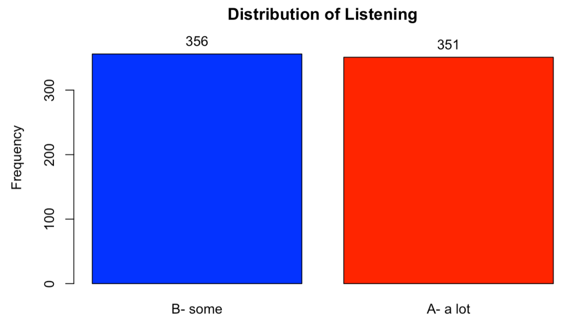
	Frequency	Percent	Cum. percent
A- a lot	407	52.4	52.4
B- some	247	31.8	84.3
NULL	61	7.9	92.1
C- not much	53	6.8	99.0
D- none	6	0.8	99.7
Response	1	0.1	99.9
C	1	0.1	100.0
Total	776	100.0	100.0

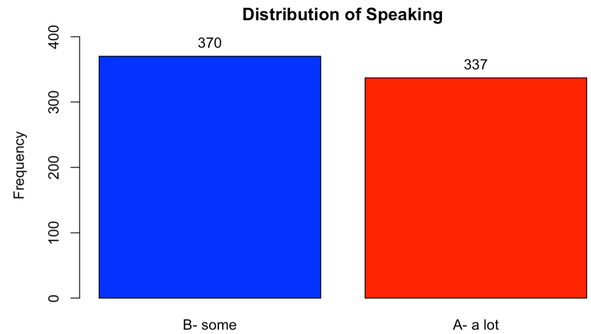
Exploratory Data Analysis: Insights

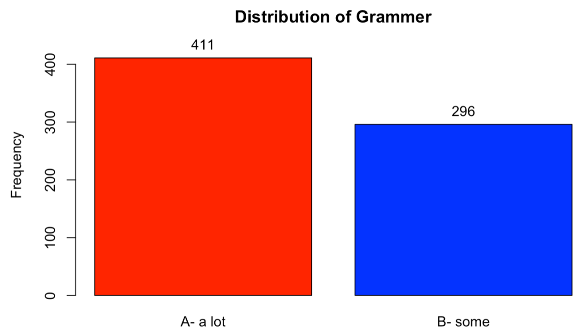
The above Exploratory Data Analysis reveals the following:

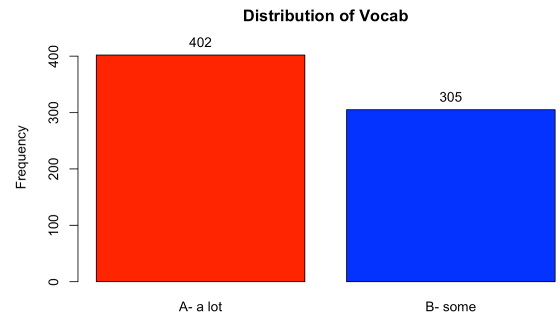
- For the variables *Listening*, *Speaking*, *Grammar*, *Vocabulary*, *Culture*, and *Recommend*:
    - Response category A had the highest frequency followed by response category B.
    - all other response categories have very low frequencies, causing significant imbalances in these variables.

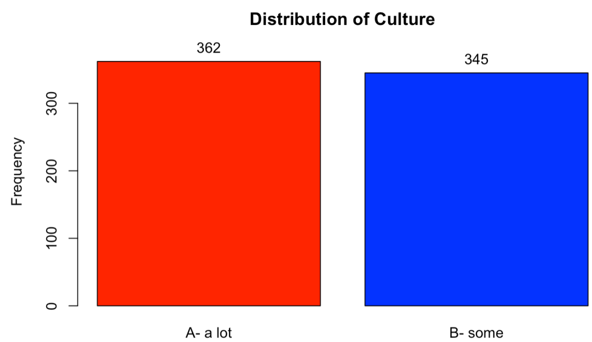
Since all of these variables had a huge imbalance within themselves, we had to transform the variables to make sure that our future models and results would be valid.
In order to rectify the imbalance, we transformed all the four predictor variables(Grammar, Speaking, Listening and Vocabulary) into binary variables. A(a lot) remained the first level, and we grouped all other levels (B-D) together to form the new level B(some).

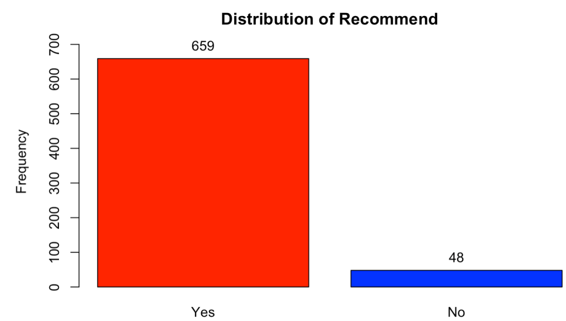










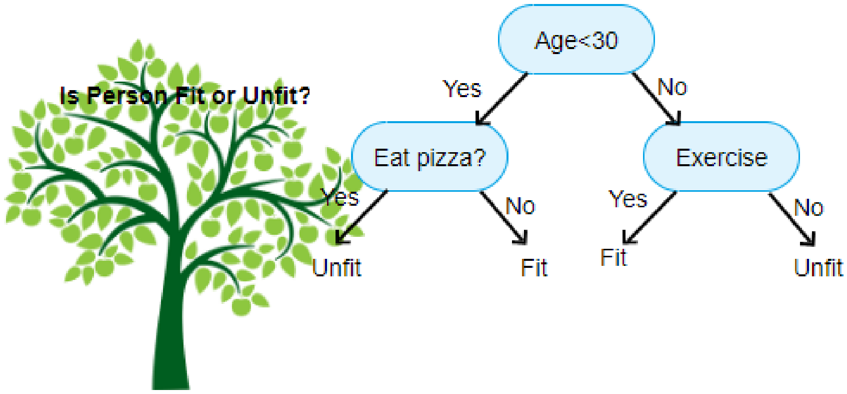


STATISTICAL ANALYSIS

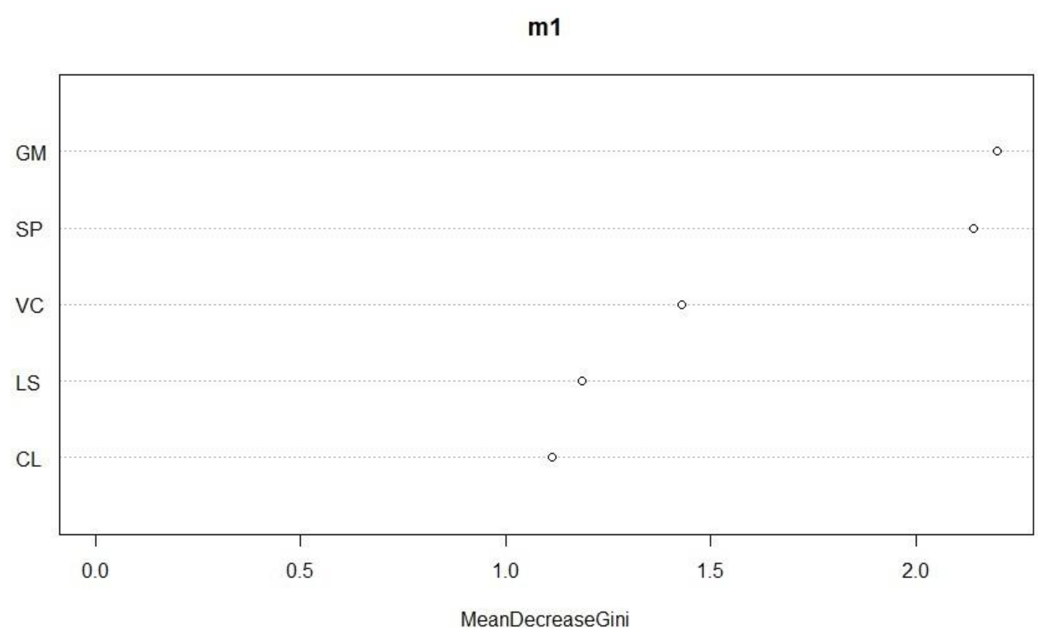
Having cleaned the data for missing and repeated values, we wanted to analyze how the students' perceived knowledge improvement in various areas like vocabulary, grammar, speaking, listening, and understanding of the American culture affected the chances of recommending the course to other people.
For this purpose, we created a data set with 6 variables which included the unique student ID, the student responses to perceived skill improvement in the above mentioned areas (On a scale of A-D, A being the highest), and the recommend/do not recommend a response.
We used two techniques to test which of these areas affected the chances to recommend a course. For both of these models, the ID served as the identifier, the skill improvement variables the predictors, and the recommendation response as the response.

Random Forest

The first technique we used was the random forest model. In order to understand this technique, we start with decision trees. A decision tree is a recursive set of conditions that allows us to reach an output. For instance, let us say we were to decide if a person is fit or unfit. As we can see in the image below, how the decision will be made depends on a number of factors including a person's age, eating, and exercise habits. Each condition is known as a node. The higher up the node in the tree, the more important that factor is in decision making.



When many of these trees are created together and then fed different data points from the same data population, each tree comes up with its own node conditions. For each observation, each tree gives its own answer, and the most common answer is chosen as the final decision for that observation.
In our case, the response or the target variable is the recommend/do not recommend student response, and the nodes are based on the perceived improvement in the five areas: vocabulary(VC), grammar(GM), speaking(SP), listening(LS), and understanding of the American culture(CL).
Using our random forest model, we get the following response:



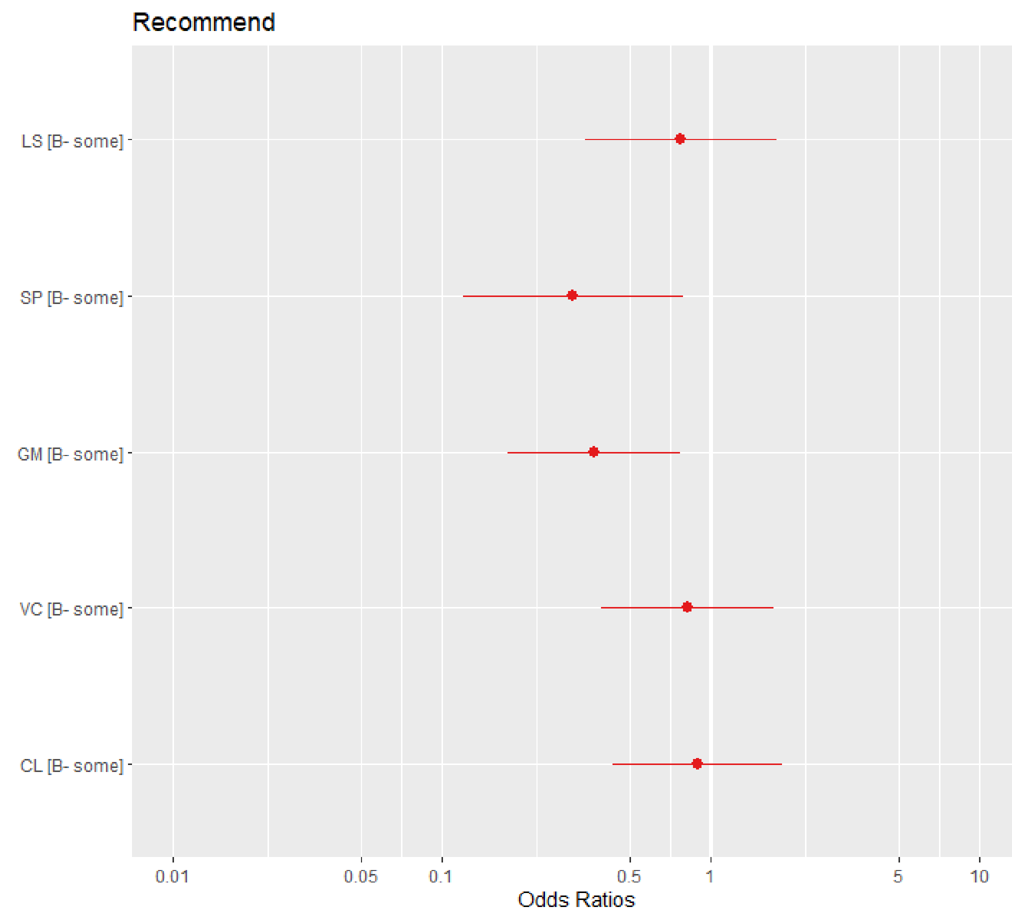
As we stated earlier, the higher up the node, the more important is that variable. As we can see here the most important node here is the Grammar node followed by the Speaking node.
The other variables do have some importance, but their value is rather insignificant.

Logistic Regression

In order to validate and check our results from the random forest model, we use the Logistic Regression model. The logistic regression model takes into account the odds ratio of an event happening and converts it into a probability. In our case, the model outputs a probability of the student recommending the course. Here are the results of this model:

Predictor	**Odds Ratio**	**95% CI**	**P-Value**
LS-B some	0.7744893	2.5% **0.3293099** 97.5% **1.7081655**	0.54035
SP-B some	0.3069496	2.5% **0.1118412** 97.5% **0.7503601**	0.01397 *
GM-B some	0.3671754	2.5% **0.1689402** 97.5% **0.7516590**	0.00809 ** .
VC-B some	0.8156880	2.5% **0.3809715** 97.5% **1.6818639**	0.58889
CL-B some	0.8961220	2.5% **0.4220043** 97.5% **1.8238769**	0.76770

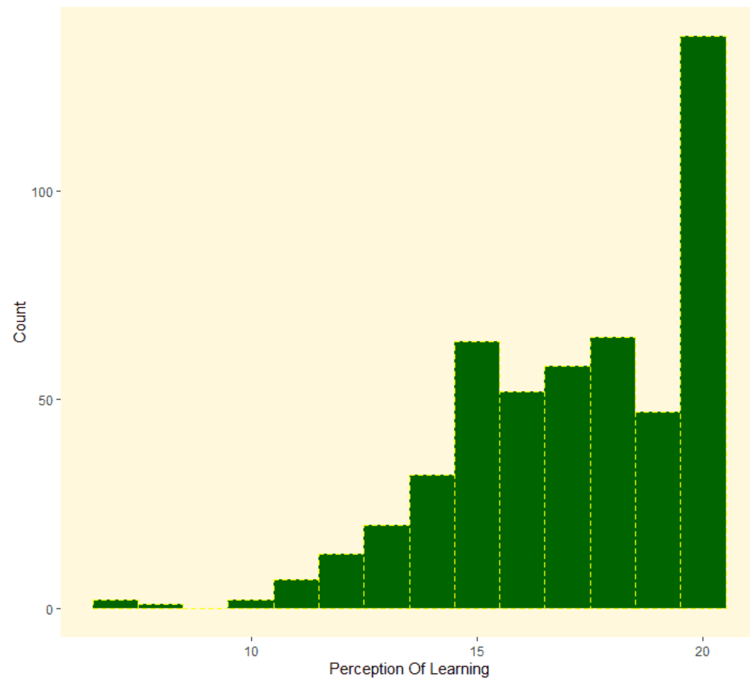
The main measure we want to focus on in this model is the odds ratio itself.
The odds ratio tells how a change in one of the predictor variables affects the response variable
The smaller the p-value in the above table, the more impact the particular variable has on the response variable.
The two predictors which have a significant impact on the recommendation are the Speaking and Grammar predictors, which is in accordance with the random forest model.
.



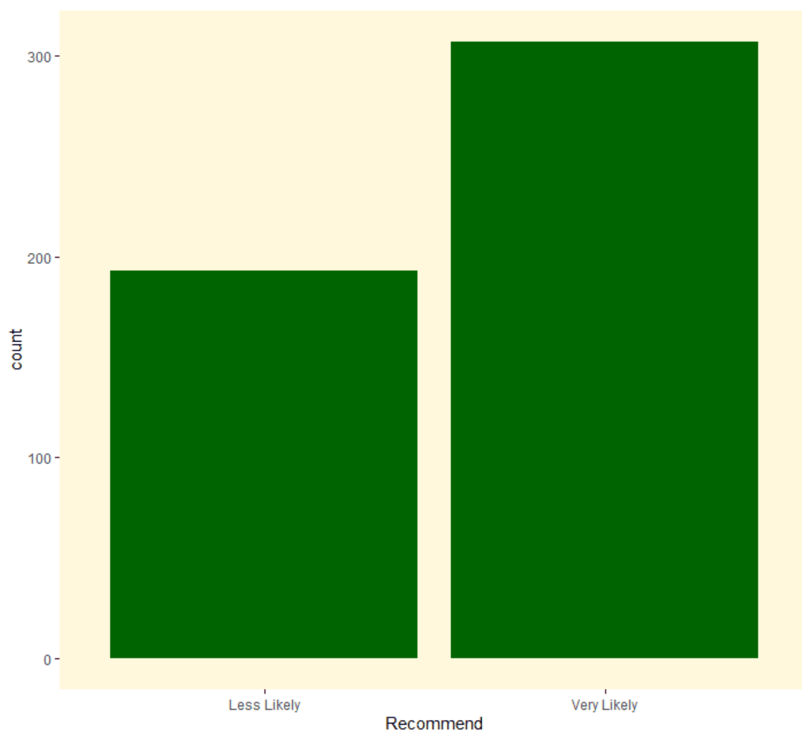
From the plot above and the table of results, we can conclude that a person who thinks that their knowledge about speaking has somewhat improved(B) as opposed to improved a lot(A) is 70% less likely to recommend the course. Similarly, a person who thinks that their knowledge about grammar has somewhat improved(B) as opposed to improved a lot(A) is 64% less likely to recommend the course.

Moving Beyond the Binary

Once we established that the perception of improvement in the above mentioned areas were significant in determining the recommendation response for students', we had grounds for moving beyond the binary recommendation/ do not recommend response.
We transformed the perceived increase in knowledge variable from a scale odf A-D(categorical) to numeric(1-4) and added them together to form a new variable "**Perception of Learning**" which had a scale of 1-16



Since we wanted to eventually categorise students' into more likely to recommend and less likely to recommend, we split this new variable using the median(17) as the threshold. Our final variable hence was a binary variable with values above 17 binned into "More Likely" and less than 17 into "Less Likely" to recommend. We named the variable "Recommend"

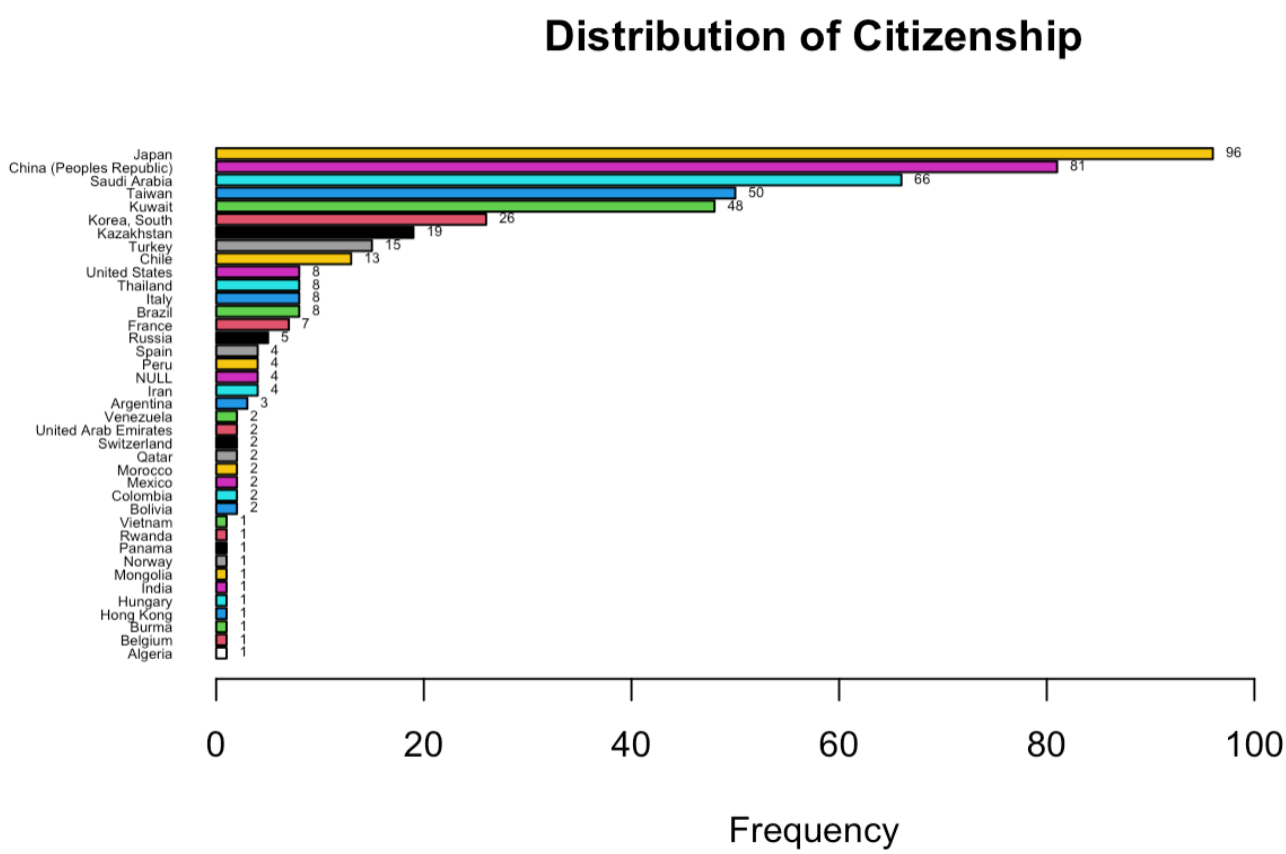


Once we had our new variable, we had to look into what influences the chances to recommend the course or alternatively, the perception of learning of the English language.
For this analysis, we wanted to focus on 3 broad categories:

- Demographic Factors - Country of origin, Age, Gender
- Teacher and Teaching Methodology Evaluation + Time Spent at UCLA extension
- Personal Life - Living situation, source of funding, how did you hear

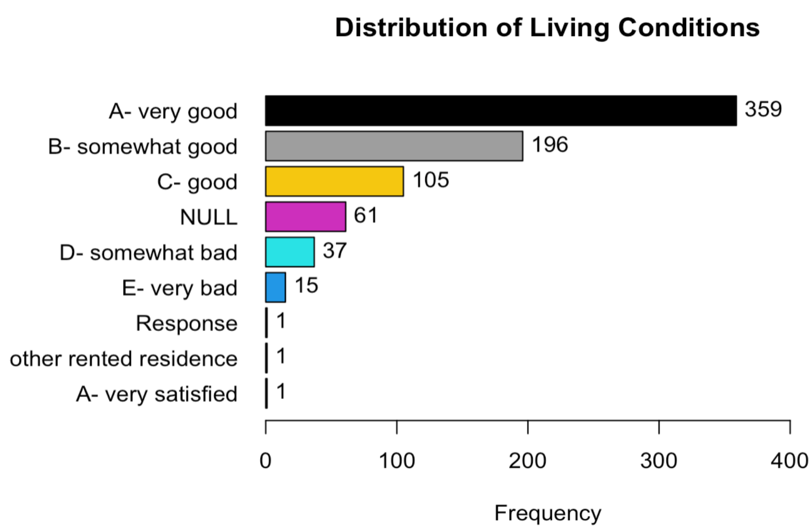
Final Analysis
For our final model, we used 9 variables -
Gender, Age, Living, Citizenship_U, Source_of_Funding, How_did_you_hear, Status
While most of the variables are self explanatory, the status here tells whether a student is a new student or a returning student. We also developed an algorithm to specify the number of times the student has taken the survey using the date of survey as the basis, but because 668 of the 707 students were first time survey takers, had a very unbalanced variable and hence we didn't include the variable in the final mode.
Moreover, survey responses for 180 students in the Spring 2016 quarter, 20 students in the Summer 2016 quarter and 3 students in the Winter 2017 quarter were missing thus reducing the data set size from 707 to 504 students.

Before we move on to the analysis, we wanted to check if our variables were balanced and ready for the model.



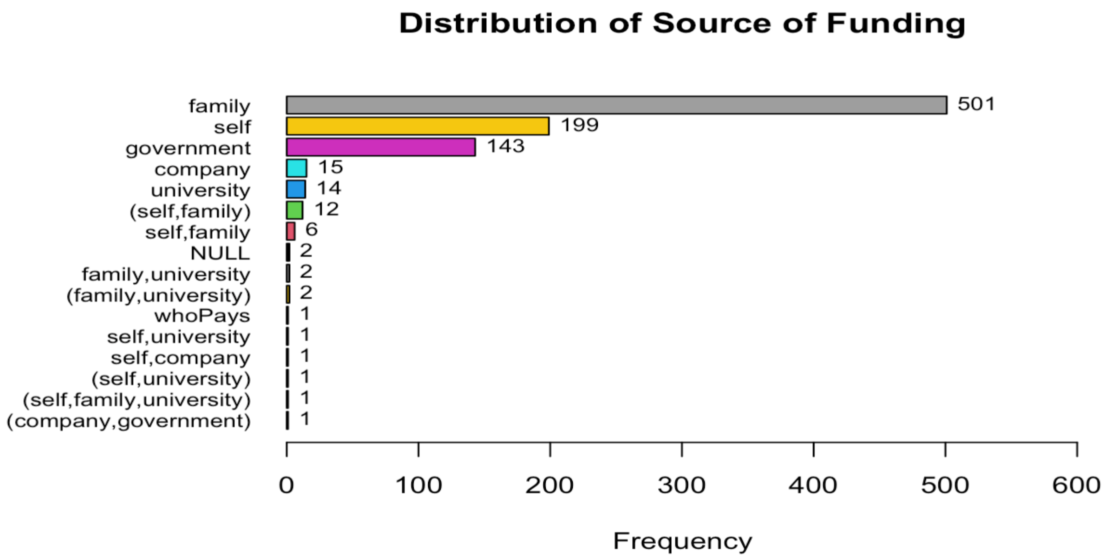
Citizenship Frequency Table

	Frequency	Percent	Cum. percent
Japan	96	19.0	19.0
China (Peoples Republic)	81	16.1	35.1
Saudi Arabia	66	13.1	48.2
Taiwan	50	9.9	58.1
Kuwait	48	9.5	67.7
Korea, South	26	5.2	72.8



Living Conditions Frequency Table

	Frequency	Percent	Cum. percent
A- very good	359	46.3	46.3
B- somewhat good	196	25.3	71.5
C- good	105	13.5	85.1
NULL	61	7.9	92.9
D- somewhat bad	37	4.8	97.7
E- very bad	15	1.9	99.6
Response	1	0.1	99.7
other rented residence	1	0.1	99.9
A- very satisfied	1	0.1	100.0
Total	776	100.0	100.0



Souce of Funding Frequency Table

	Frequency	Percent	Cum. percent
family	501	55.5	55.5
self	199	22.1	77.6
government	143	15.9	93.5
company	15	1.7	95.1
university	14	1.6	96.7

Note: Frequency Tables for *Citizenship* and *Source of Funding* only show a portion of the complete table.

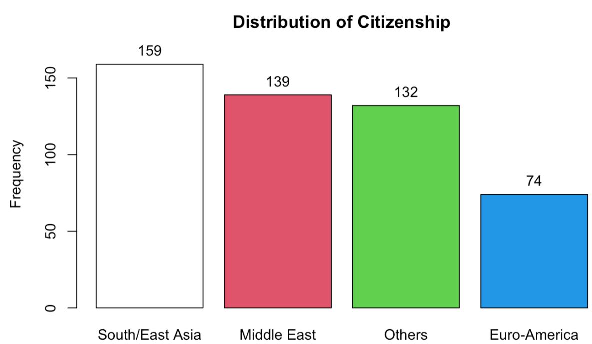
- For the variable *Citizenship*:
    - There are 39 countries, and the top 6 countries from which students come from are Japan, China, Saudi Arabia, Taiwan, Kuwait, and South Korea.
    - The other bottom 29 countries have significantly lower frequencies than the top 10 counties.
    - The variable citizenship shows significant imbalance as there are many countries that have very low frequencies.
- For the variable *Living Condition*:
    - 85.1% of students claimed that their living conditions were either very good, somewhat good, or good.

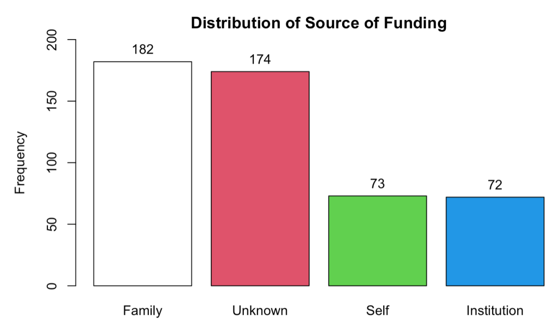
- 46.3% claimed that their living conditions were very good.
- 6.7% claimed that their living conditions were somewhat bad or very bad.
- This creates an imbalance as there are only 6.7% of students who claim that their living conditions were somewhat bad or very bad.

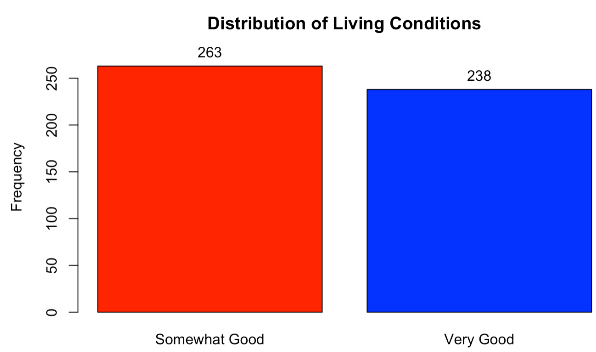
- For the variable *Source of Funding*:
  - 55.5% of students claimed to have been funded by family
  - 22.1% of students claimed to have been self-funded
  - 15.9% of students claimed to have been funded by government
  - 93.5% of students claimed to have been funded by family, by themselves, or their governments while 6.5% of students were were funded by a company, university, or combinations of all sources of funding
  - This creates an imbalance where several categories have very low frequencies

Since all of these variables were unbalanced within, we had to transform them to ensure our future analysis was valid.

Transformed Data







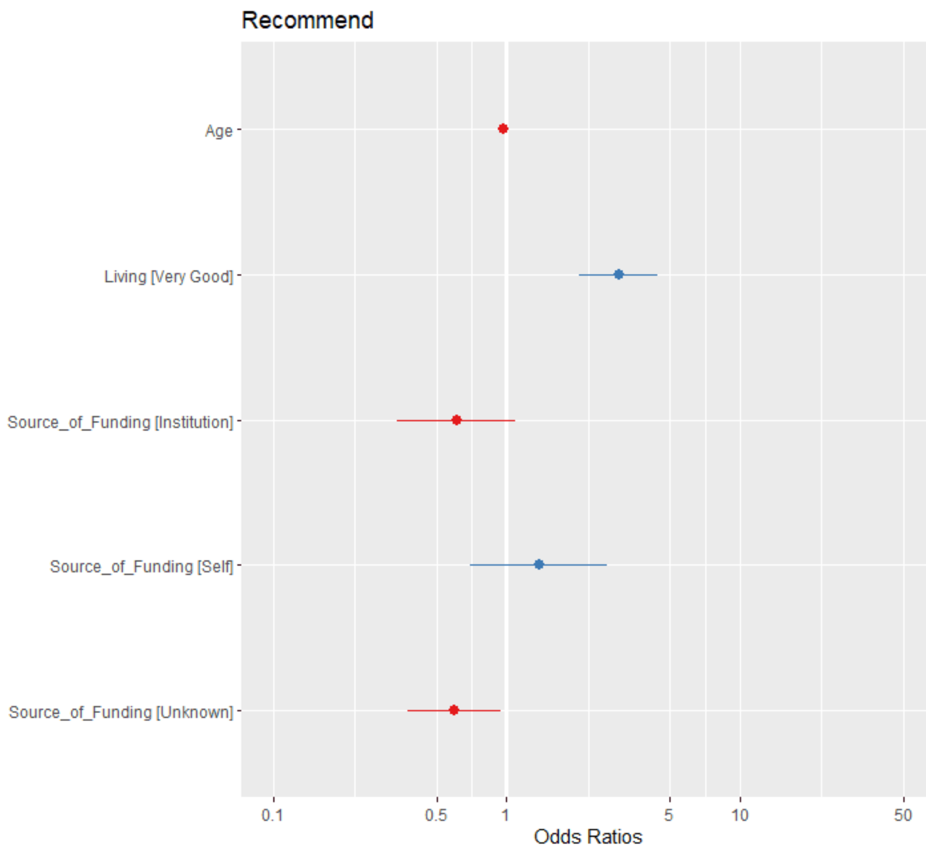
For the citizenship variables, we divided the data into 4 major categories, South/East Asia, Middle East, Euro-America(Containing all countries in Europe and America) and Others. As for the source of funding, the combination responses like self/company, were evenly distributed among the major categories: Family, Self, Institution (Company + University + Government).

Lastly, just like we made a binary variables out of our Grammar, Vocabulary, Culture, Listening and Speaking Variables, the living conditions variable was divided into two categories: A (Very Good) and B(Somewhat Good)

Once we transformed all the variables to be suitable for model building, we used the logistic regression model explained earlier to find out the relationship between the variables we chose and the perceived level of learning/likeliness to recommend.
Since there were a huge number of predictors, we used backward stepwise logistic regression for variable selection. In this technique, the algorithm takes the full model with all predictors and systematically removes the unnecessary variables which don't add predictive power to the analysis. Using this technique, we were able to narrow our number of predictor variables to 3:
Age, Living Condition and Source of Funding.
Our final model gives the following results:

Predictor	**Odds Ratio**	**95% CI**	**P-Value**
Age	0.9664169	2.5% **0.9377147** 97.5% **0.9955524**	0.02464 *
LivingVery Good	3.0258495	2.5% **2.0632053** 97.5% **4.4789852**	2.08e-08 ***
Source_of_FundingInstitution	0.6086909	2.5% **0.3394709** 97.5% **1.0935917**	0.09538 .
Source_of_FundingSelf	1.3764404	2.5% **0.7068939** 97.5% **2.7366441**	0.35340
Source_of_FundingUnknown	0.5923392	2.5% **0.3726636** 97.5% **0.9366424**	0.02572 *

Here, we can see that with every 1 year increase in the age of the student, the perception of learning goes down by 3.8% and hence they are 3.8% less likely to recommend the course. The most significant predictor was the Living Condition, where a person who reported their living condition as very good as opposed to somewhat good, was 3 times more likely to have higher perception of learning as well as chances to recommend the course. As for the funding, the odds ratio here are in reference to funding from the family. So a person who pays for the education theirself, is 37% more likely to have a greater perception of their learning improvements and thus would be more likely to recommend the course.



Using this logistic regression model, we got an accuracy of 61.6% for the new response variable.

Conclusions

In this research, we investigated whether we could find better measures of the ALC program's success, beyond the binary recommend/don't recommend response. After organizing/cleaning the data and exploratory data analysis, we decided to create a new variable of focus: 'perception of learning'. First, we summed up the variables of interest for each student, 'grammar', 'speaking', 'vocabulary', 'listening', and 'culture', after changing the values from a categorical A-D scale to a numerical 1-4 scale. Using the median from this collection of sums, the sums were sorted into 'Less likely' or 'More likely' depending on whether the sum was lesser or greater than the median. Our final model consists of three variables: 'Age', 'Living Conditions', and 'Source of funding'; these variables were most important in predicting the student's perception of learning from the ALC program. The model had a 61.6% accuracy using these three variables and was able to correctly identify 24 out of the 29 people who wouldn't recommend the ALC program. Due to the unproportional sample size, our research does not claim any statistical significance, but our research does show that a student is more likely to recommend the ALC program if they had a positive perception of their learning from the program. It is important to note that after the data cleaning required for considering the additional new variables, we ended with 500 data points from the original 776 observations.

As a result from this research, we suggest that reducing the options in the survey from a A-D four-choice scale to a A-B binary scale would be beneficial, as the less positive options C and D are rarely chosen. Also, we also suggest restricting the responses to predetermined choices rather than requiring a free response; this action would likely increase the response rates. For further research, we would like to recommend the use of the results from this research in creating a new survey and then proceeding on analysis from the data of the new survey; for example, since older students were less likely to recommend the ALC program, it may be ideal to create a separate, more detailed survey for older students to have a better idea on why they may not recommend the program.

Shortcomings

Although our analysis provides some crucial information regarding the student satisfaction for the courses at the American Language Center, there are a few shortcomings that need to be addressed.
The initial dataset which we chose to analyze the improvement variables had a very skewed distribution for the response. Out of the 707 total observations, 658 students said they would recommend the course and only 48 students said they would refrain from recommending the course to other students.
Once we transformed our initial recommendation variable to the perception of learning variable, we had to get rid of a few more observations because of incompleteness of data and were left with 504 observations where only 29 people chose to not recommend the course.

Here is a table of comparison between our old response(No/Yes) and our new response(Less/very Likely)

	Less Likely	Very Likely
No	24	5
Yes	169	302

Based on the comparison of the old and the new variable

- The odds of recommending the course being in the Less Likely category is 7.04 while the odds of recommending the course in the very likely section is 60.
- The chances of recommending the course in the very likely section is ~8.6 times when compared to the less likely section
- Our new metric correctly captures 24 out of the 29 people who won't recommend this course

Although the response variable is very skewed and our new metric isn't perfect, it still manages correctly captures 24 out of the 29 people who won't recommend this course. With more data for students in the coming quarters, we believe that this rate would go up.