UNIVERSITY OF CALIFORNIA

Los Angeles

Using Machine Learning to Aid
Second Language Acquisition

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics & Data Science

by

Maryam Zaki Akbar

2024

ABSTRACT OF THE THESIS

Using Machine Learning to Aid
Second Language Acquisition

by

Maryam Zaki Akbar
Master of Applied Statistics & Data Science
University of California, Los Angeles, 2024
Professor Yingnian Wu, Chair

Adults are often told that it may be too late for them to learn a new language, and, sometimes, they may even be intimidated out of continuing their learning journey. While it may be a daunting task, it is not an impossible one. In this thesis, using Duolingo's dataset of about 13 million learning traces, four different machine learning models are fitted to predict whether the probability of recall of a word is greater than or equal to 0.5. Of the four, logistic regression fared the best with an accuracy score of 93%. It also identified certain word features that contribute to improving the chances of recalling a word.

The thesis of Maryam Zaki Akbar is approved.

Oscar H. Madrid Padilla

Frederic R. Paik Schoenberg

Maria Cha

Yingnian Wu, Committee Chair

University of California, Los Angeles

2024

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

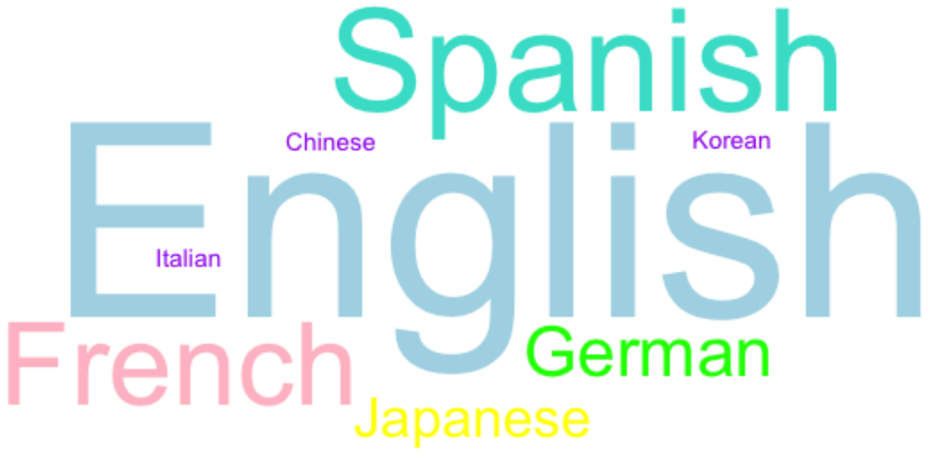


Figure 1.1: Languages Learned Sized by the Number of Countries They Are Most Popular in (Blanco, 2023)

English is the most popular language being learned in 122 countries as can be seen by the above figure (Blanco, 2023). It has long been thought that many people who try to learn a new language as an adult struggle to achieve native proficiency, in fact, research has found that they can outperform in proficiency in comparison to younger native speakers, while mature native proficiency as an adult is limited due to exposure time required (Hartshorne, Tenenbaum & Pinker, 2018). Dabrowska & Street posit that clearly teaching the language allows graduate and non-graduate non-native speakers to develop better metalinguistic ability and perform better than non-graduate native groups when tested on their second language (2006). Regardless of level of education, the group is native and non-native speakers' ability to surpass them in this respect signals that motivation and access to classroom instruction can aid in acquiring a second language. In later works, Dabrowska found that 75% of non-native learners performed similarly to native speakers when tested on grammar and 94% performed similarly when tested on vocabulary (2018).

While second language acquisition (commonly known as L2) as an adult may be perceived as challenging compared to learning as a child, it is worth noting that online language learning has advanced considerably. This varies from active learning through language learning platforms like Duolingo, launched in 2011, and Babbel, launched in 2007, to relatively passive learning through language exposure in a more globalized world (Duolingo, n.d; Babbel, n.d). Social media users can translate captions in foreign languages to their native language while website viewers can translate entire websites in mere seconds. This poses the question how much does passive exposure contribute to second language immersion.

While there are various ways to learn new languages, how does one ensure stickiness in second language acquisition, i.e., how does one retain newly acquired vocabulary and grammatical knowledge? It may be easy to develop a foundation by learning commonly used words and phrases, learn pronunciations of the L2 alphabet, and even understand grammatical rules. However, how long does this information last in one's memory and what can be done to elongate its time within memory and therefore, develop long term knowledge, even close to native level knowledge? One great way would be learning through flashcards (Zarrati, et al, 2024). Additionally, for L2 learners, do certain word features contribute to identifying the word?

This thesis will focus on active learning through language learning platforms, in particular, Duolingo. It will assess different machine learning models and different word features, like whether it is a verb, singular, among others, and exposure features, like how often the word is seen in the past and use them to predict if the individual will be able to recall the word or not.

CHAPTER 2

Literature Review

2.1 Original Paper

Settles and Meeder (2016) introduced half life regression to understand how fast a word decays using Duolingo's data. Using Ebbinghaus' (1885/1913) study on forgetting, they claim that a word introduced after d or *delta*, the time since the word was last seen, surpasses the half-life (h) of the word in a learner's memory, the word may no longer be recalled. Therefore, to increase probability of recall (p) of word in memory, the word should be presented again to the learner:

$$p = 2^{-d/h} \tag{2.1}$$

They calculate half-life as seen in Equation 2.2 based on the assumption that half-life increases exponentially every time a learner sees a word. Their half-life regression model was able to reduce errors in probability of recall predictions. In the equation below, Θ is the parameter vector and $\mathbf{x}$ is the feature vector for each learning trace.

$$\hat{h_\Theta} = 2^{\Theta \cdot \mathbf{x}} \tag{2.2}$$

2.2 Other Research

Hartshorne, Tenenbaum & Pinker (2018) implemented an Exponential Learning with Sigmoidal Decay model to predict grammatical proficiency, g, using age when exposure began, t_e, age when they were quizzed, t, learning rate, r, and an experience discount factor, E:

$$g(t) = 1 - e^x \tag{2.3}$$

Here x is the following:

$$x = \int_{t_e}^{t} -Erdt \tag{2.4}$$

The authors conducted their study by collecting responses online through a quiz to make sure they had a large audience. The discount factor that they incorporated was specific to whether the person responding to the quiz had experience with the language. This experience ranged from bilingual, immigrants and non-immersion learning individuals. The authors were able to find in their study that adults tend to outperform children in second language acquisition. Interestingly, they found that for an individual to be as proficient in a language as a native, they would have to be exposed to the language before the age of seventeen.

Seibert Hanson & Brown (2019) conducted an experiment with 62 Spanish learners using Anki, a flashcard app for language learning. They first gathered information from the learners to determine the motivation, language efficacy, baseline Spanish knowledge among other metrics. Students were then given access to Anki as well as told how to use it. At the end of the experiment they were asked about the same metrics as at the beginning of the experiment and also asked about their study styles for their Spanish class in particular and subsequently tested on their Spanish. Through their research they found that spaced repetition helps in language learning. However, coupling it with other kinds of learning strategies enhances the language learning process. Additionally, motivation is crucial in language learning.

CHAPTER 3

Dataset

3.1 Background

In my search for how second language acquisition has evolved, I came across this dataset on the Harvard Dataverse (Settles, 2017). The dataset is from Duolingo and was used in their research (Settles & Meeder, 2016). This dataset has approximately 13 million learning traces and twelve variables. Each learning trace has the **timestamp** the student begins the session, the **user_id**, and the word, or **lexeme_id**, which makes the learning trace unique. Additionally, it identifies what the language being learned is, **learning_language**, as well as what language the interface is in, **ui_language**. Next, it has metrics to gauge **p_recall**: the student's understanding of the word by calculating the probability the student recalls the word, **delta**: the time passed since that word was seen, **lexeme_string**: details regarding the word, **history_seen** and **history_correct**: history of exposure to the word in previous sessions and the number of times it was identified correctly and, lastly, **session_seen** and **session_correct**: the results of the current session.

With regards to details of the word, the **lexeme_string** variable holds information regarding the form the word was seen in, the root, the part of speech, and any relevant modifiers. The dataset has words with various different types of parts of speech. These include the categories: nouns, adjectives, verbs, adverbs, conjunctions, determiners, interjections, numerals, prepositions, among others. A word can have multiple modifiers as these give more information regarding how the word may have been used in the exercise. The modifiers in the dataset detail the tense of the word, if it is singular or plural, its gender, if it was used in the first person, second, or third person. An example of a lexeme string for the German word *lernt* would be as the word is presented in the exercise, its root *lernen*, the fact that it is a verb, its tense being present indicative, that it is used in the third person, and that it is used as a singular verb.

3.2 Exploratory Data Analysis

3.2.1 Learner Demographics

As we can see that of all languages English followed by Spanish are the largest languages being learned on the Duolingo platform. Similarly, English and Spanish also tend to be the largest native or user interface languages. In fact, nearly 40% of all learners in the dataset are learning English, while approximately 60% of all learners use English as their user interface language.

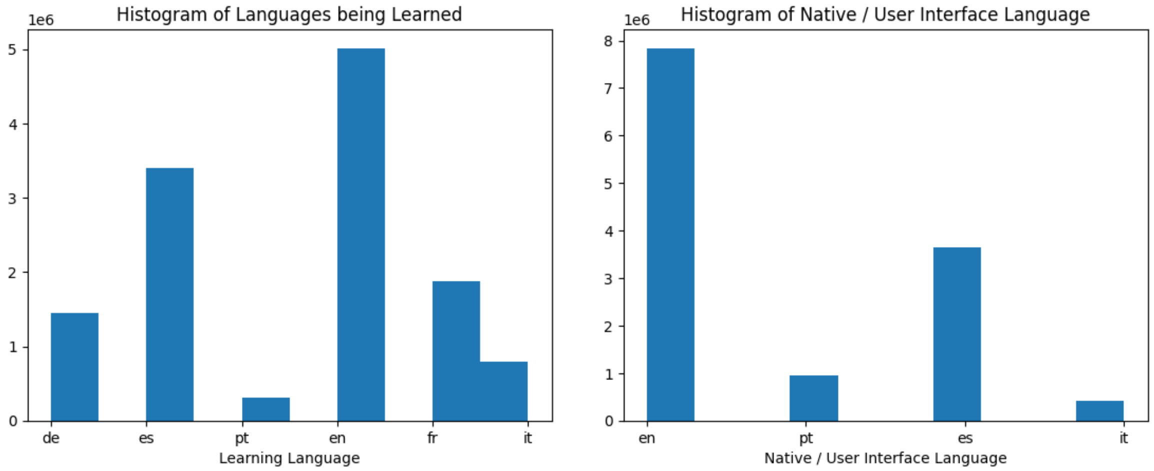


Figure 3.1: Languages being learned vs. User Interface Language

While, it's tempting to assume given this information that most users come from English speaking countries, it may not be the case and many may be using English as an interface language as their native language may not be available as an option.

3.2.2 Learner Usage

Table 3.1: Snapshot of Learners' Usage

Summary of Learners	
Count	115,222
Max Number of Sessions / Learner	19,194
Min Number of Sessions / Learner	1
Average Number of Sessions / Learner	112

The dataset has information on 115,222 learners on the Duolingo platform, who on average have had 112 sessions for the period of time this data was recorded. Some learners seem to be more committed than others with the maximum number of sessions per learner being 19,194 and the minimum only being 1. While these learners seem to be outliers since the average session is 112, it may be of interest to investigate why these learners are acting in such a way. Is it because they are personally motivated or is the app trying to increase engagement?

Additionally, we know that the dataset contains sessions starting Feb 2013 and ending in March of the same year.

Summary of Delta	
Max Amount of Time	466 days, 65962 seconds
Average Amount of Time	8 days, 3831 seconds, 61804 microseconds
Standard Deviation of Amount of Time	26 days, 99 seconds, 14600 microseconds

Table 3.2: Snapshot of When the Word was Last Seen

It can also be seen that learners are not that sticky in terms of their routine visits to the app for second language learning. However, it may be the case that the average learner has a weekly routine to check into the app to continue their learning. A deep dive into delta later in this chapter can highlight how it changes across the learning traces within the dataset. Additionally, there are some metrics we do not know in terms of these learners' time on the app. Some learners may have been on the app for a while and either developed a routine or only check in a few times to keep up with their progress.

History of recall is a variable created by dividing history of correctness by history of exposure to the word. A comparison of the summary statistics for probability of recall and history of recall indicates that the mean of history of recall is slightly higher

Summary Statistics	**Probability of Recall**	**History of Recall**
Max	1.000	1.000
Min	0.000	0.045
Mean	0.896	0.901
Standard Deviation	0.271	0.136

Table 3.3: Summary Statistics of Recall Metrics

than probability of recall. There is also more variance in probability of recall compared to history of recall. This difference in the variances between the two metrics may be because of delta. The relationship between all of these variables will be further investigated later in this chapter.

3.2.3 Further Investigation into History of Seeing the Word

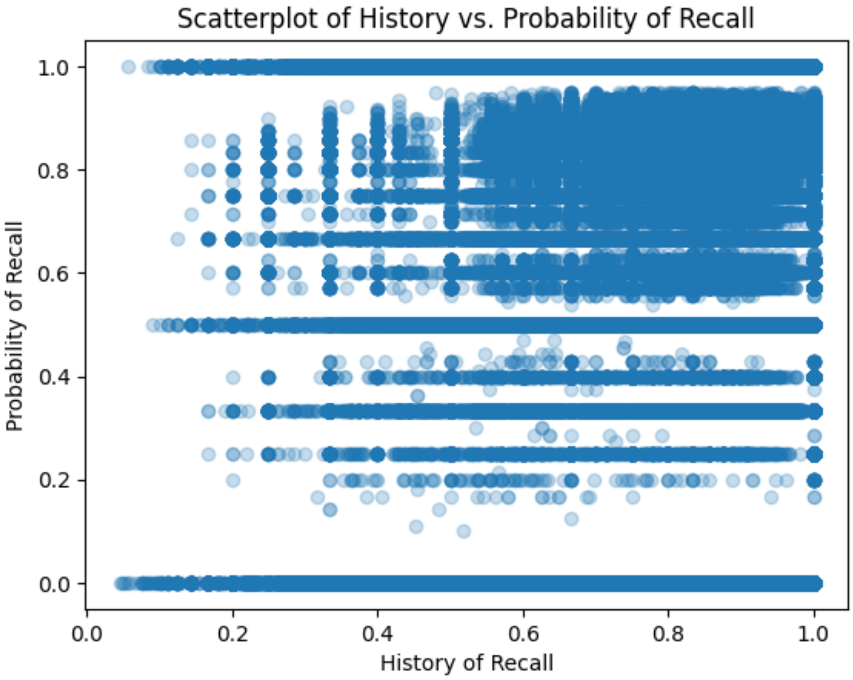


Figure 3.2: History vs. Probability of Recall

In Figure 3.2, while there does not seem to be a linear relationship between the two variables, it is apparent that for higher history of recall, a large quantity of the points for probability of recall are concentrated towards the higher end of the scale from 0 to 1.

Summary Statistics	History_Seen	History_Correct
Max	13,518	12,888
Min	1	1
Mean	21.98	19.35
Standard Deviation	129.55	111.97

Table 3.4: Summary Statistics for History_Seen and History_Correct

In the above table we can see there is quite a lot of variance in both variables, history_seen and history_correct, the mean indicates that in general learners do use the platform quite often and have developed a history interacting with words on the platform.

3.2.4 Further Investigation into Delta and a User's interaction with a certain word

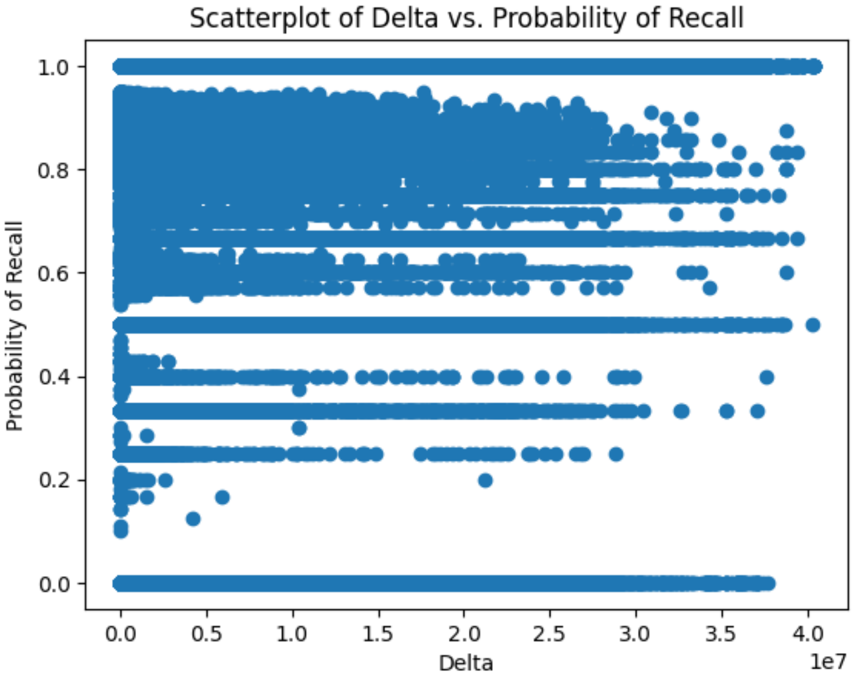


Figure 3.3: Delta vs. Probability of Recall

Similarly in Figure 3.3, while we, again, do not see a linear relationship, for lower values of delta there are larger concentrations of higher probability of recall. This is in line with Equation 2.1, as delta increases, probability of recall decreases.

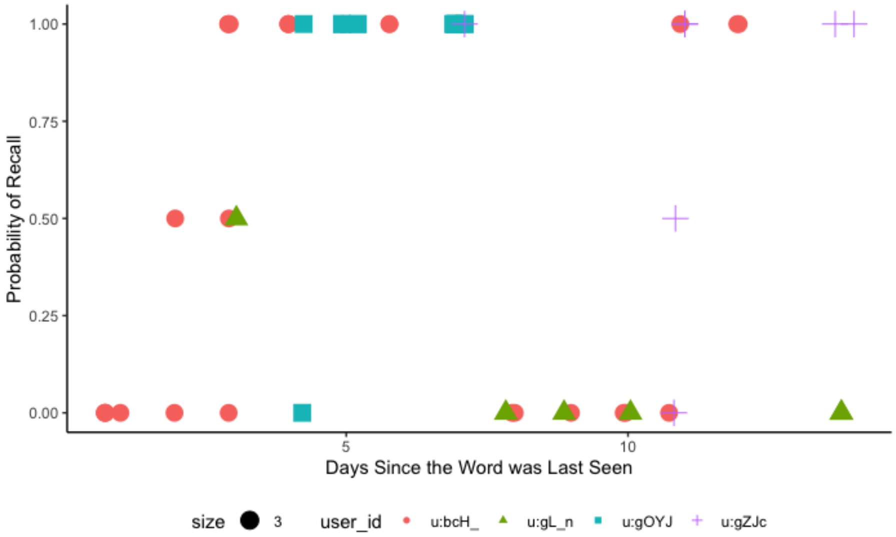


Figure 3.4: Word Learning Trace (Settles & Meeder, 2016)

Let's further hone in on four users and how their probability of recall changes over time. Here, each symbol shows the users' contact with the word. Settles & Meeder (2016) map out a similar journey in their paper for a user learning a French word. In Figure 3.4, I map the journeys of four different users learning the English word temperature. For user bcH_, they interact with the word quite a bit and are able to bring their probability of recall higher, however, they do not keep up their practice or see that word again in their practice which is why their probability of recall falls again. For user gL_n, they find that at the beginning of their journey, the user was about to forget the word. While their probability of recall was already quite low, it would soon fall under 0.5 as more time passed without them having been exposed to the word. Meanwhile, user gOYj seems to be recall the word perfectly, due to their constant exposure the word. Lastly, user gYJc also seems to be doing well in recalling the word until some more time passes and they are unable to recall it.

3.2.5 Further Investigation into the Lexeme String Variable

Part of Speech	**Lexeme Code**	**Examples in English**
Verb	*vaux / vbmod*[a] *vbdo*[b] *vbhaver*[c] *vbser*[d] *vblex*	can, would, will do, does, did has, have is, am read, eat
Noun	*n* *np*[e]	bread, boys, water England, America
Adjective	*adj*	green, new, tired
Adverb	*adv* *cnjadv*[f]	why, when, where until, while, because
Determiner	*det*	the, an, a
Interjection	*ij*	hi, hello, bye
Numeral	*num*	one, two
Preposition	*pr*	among, near
Pre-Determiner	*predet*	all
Preadverb	*preadv*	too, very
Pronoun	*pn / prpes*	we, he, whose
Coordinating Conjunction	*cnjcoo*	and, or
Subordinating Conjunction	*cnjsub*	if, whenever, though

Table 3.5: Parts of Speech

[a] Auxiliary or Modal Verb
[b] Verb to do
[c] Verb to have
[d] Verb to be
[e] Proper noun
[f] Conjuctive Adverb

Tense	**Lexeme Code**	**Surface form / Lemma**	**Language**
Conditional	*cni*	pourrais / pouvoir	French
Future Indicative	*fti*	pourra / pouvoir	French
Gerund	*ger*	cooking / cook	English
Preterite Indicative	*ifi*	compris / comprendre	French
Imperative	*imp*	mange / manger	French
Infinitive	*inf*	have / have, open / open	English
Past	*past*	said / say, saw / see	English
Imperfect Indicative	*pii*	pouvait / pouvoir	French
Imperative Subjunctive	*pis*	pudiera / poder	Spanish
Past Participle	*pp*	presented / present	English
Present Participle	*pprs*	going / go	English
Present Indicative	*pri*	is / be, eats / eat	English
Pronomial	*pron*	disculpe / disculparse	Spanish
Present Subjunctive	*prs*	puisse / pouvoir	French

Table 3.6: Tenses

The lexeme string variable may help to bring more complexity to the model as it helps assign features to the word that will predict if a user is able to recall a word or not. The part of speech the word takes up either in a sentence, audio recording, or a photo determines how well a learner is able to remember it, similar to tense and other word features. Within this chapter, Tables 3.4 and 3.5 show the different parts of speech and tenses as well as relevant examples for clarity. There are also other word features, for example, if the word is in first, second, or third person, if it is singular or plural, if it is numeric, among other such features (Settles, 2016). The lexeme string variable is broken into categorical variables with values 0 and 1 for each feature. More modifiers can be found in the Appendix.

3.2.6 More Word Features

In their research, Settles & Meeder (2016) note that word length could be an important feature as well to incorporate. Therefore, to further understand how words are perceived by learners and ultimately use these variables to predict probability of recall, I created two more features:

1. Word Length:

   This calculates how long the word in its surface form presented to learners is. As can be seen from Figure 3.5, most words are not that long. They tend to be around 5 letters.

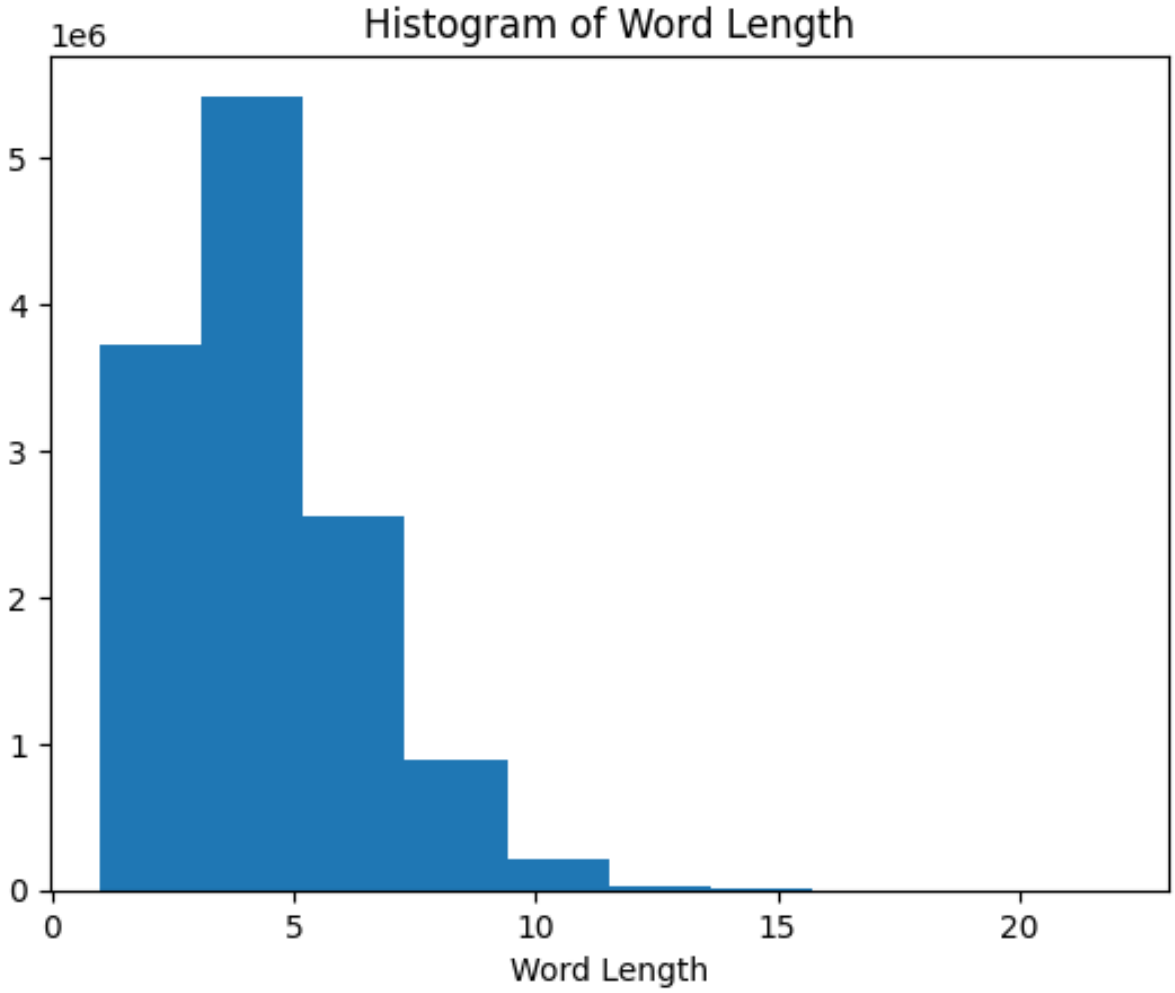


Figure 3.5: Histogram of Word Length

2. Lemma Length: This returns how long the lemma of the word is. Similarly and unsurprisingly, the lemma of most words is not very long either.

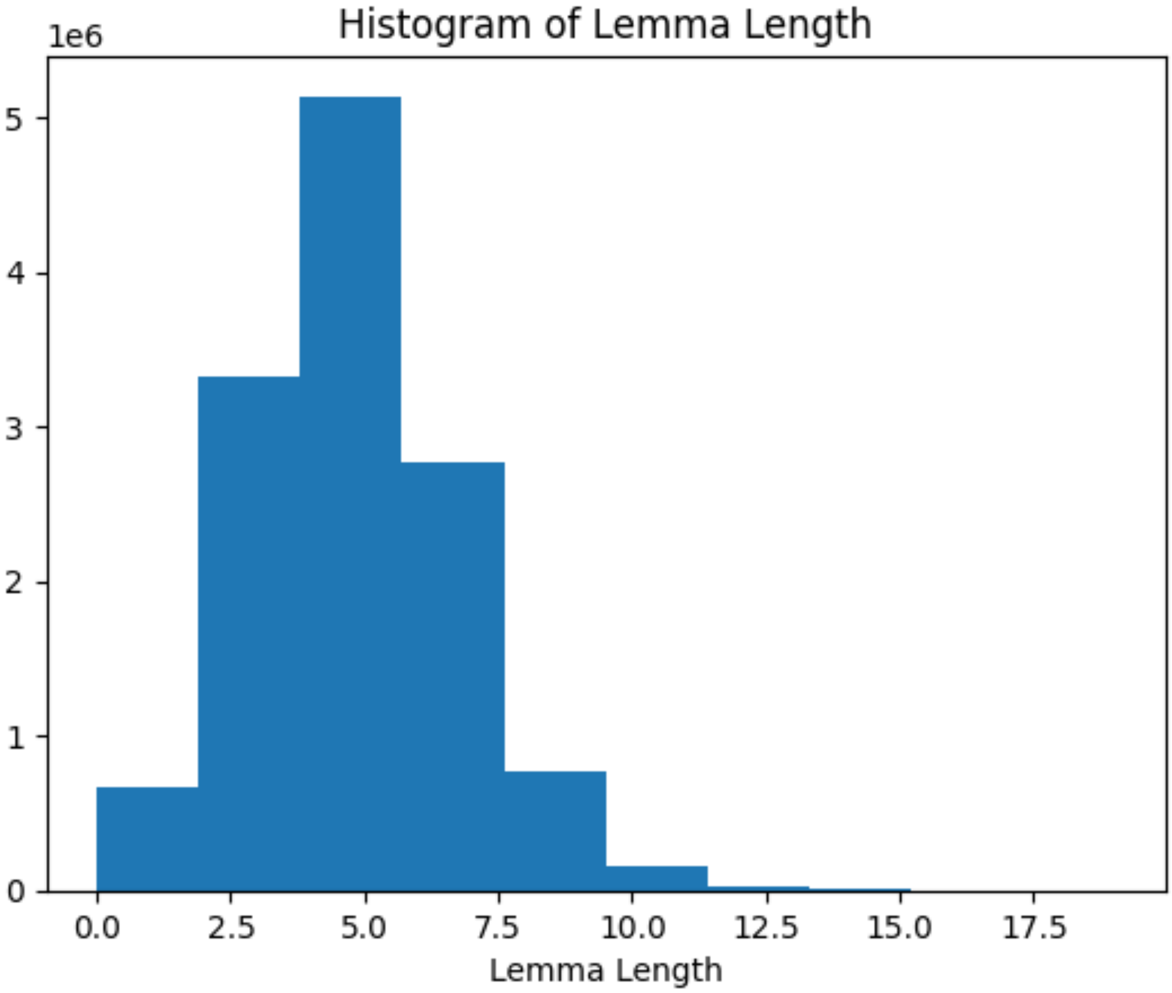


Figure 3.6: Histogram of Lemma Length

3.2.7 Probability of Recall

Probability of recall for most traces is greater than 0.5, this shows that for most of the words being learned on the platform, their half lives in memory have not passed yet.

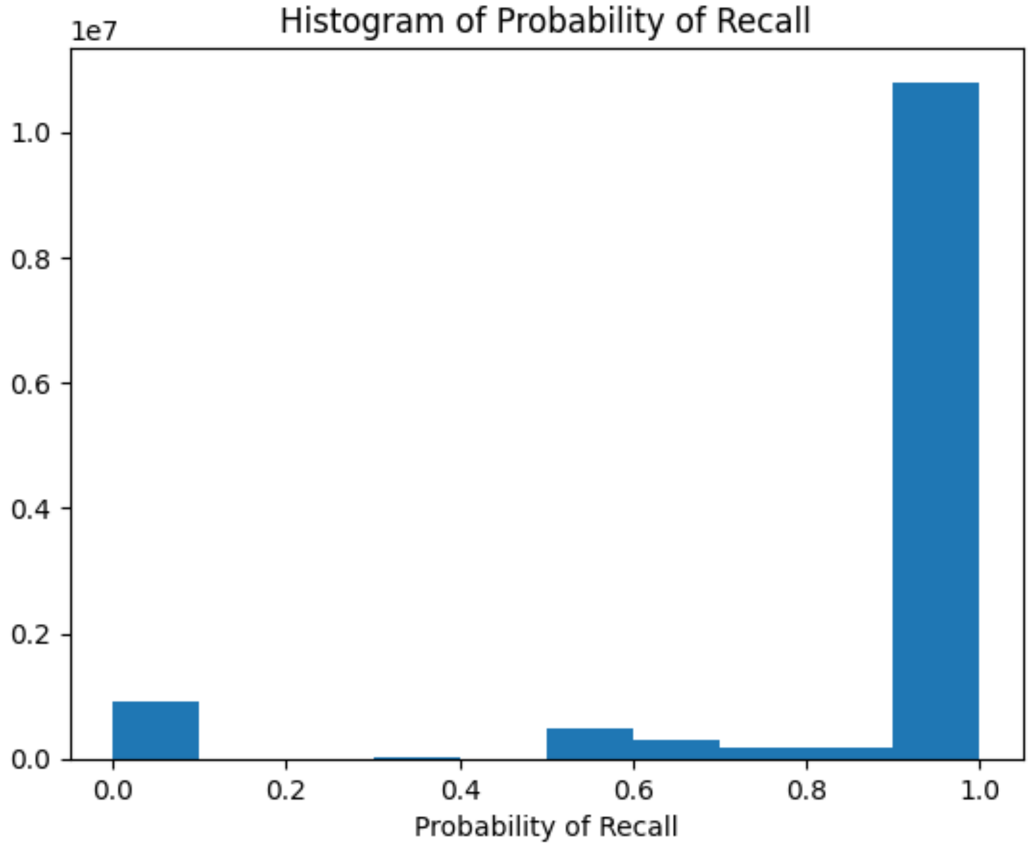


Figure 3.7: Probability of Recall (Continuous)

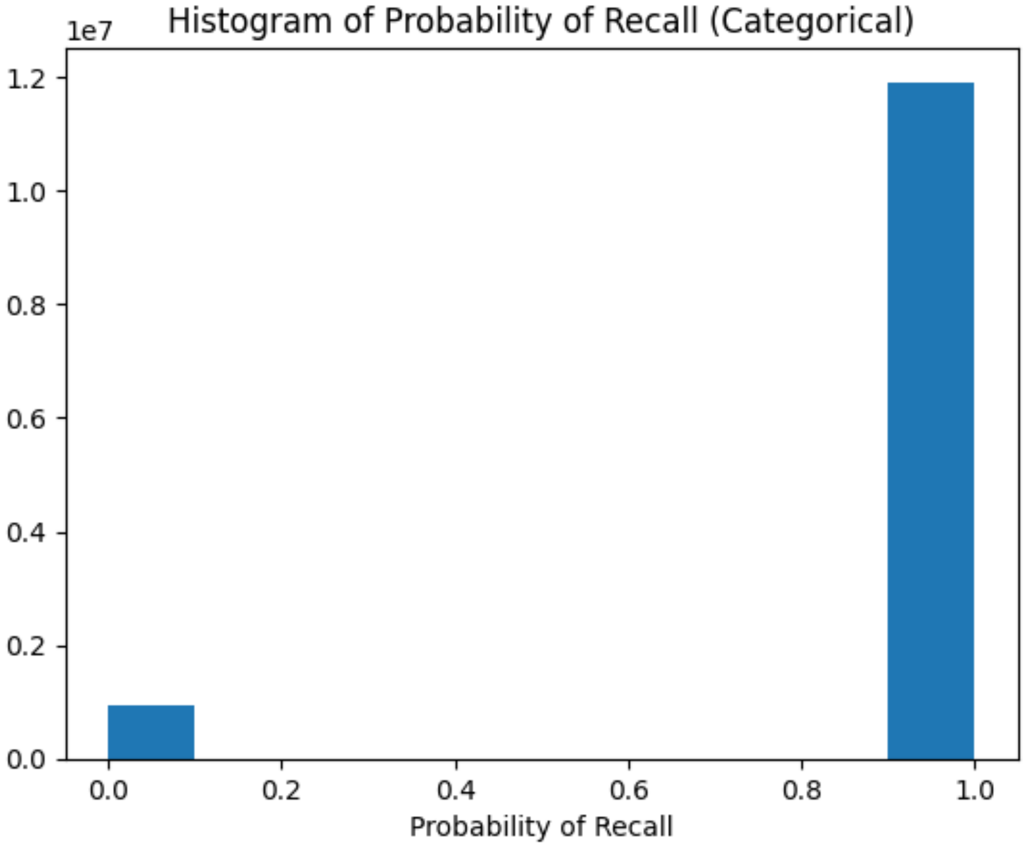


Figure 3.8: Probability of Recall (Categorical)

For ease of use, I transform the probability of recall from a continuous variable to a categorical one called p_recall_np. For probability of recall greater than or equal to 0.5, p_recall_np is 1 and otherwise 0.

CHAPTER 4

Fitting Different Models

4.1 Research Question

How can machine learning improve the probability of recall of or the ability to recall a new word in one's L2 language? And what predictors are most important? The predictor variables will be chosen from the lexeme string features (Section 3.2.5), word and lemma lengths, history of being exposed to the word in sessions, and delta, when the word was last seen. Through the following models, I aim to predict the target variable, ability to recall:

1. Logistic Regression
2. Half Life Model
3. Decision Tree Classifier
4. Random Forest Classifier

4.2 Logistic Regression

Logistic regression predicts using a logistic function (Scikit-learn, n.d.). Using this classifier, I build namely three models:

1. In this model, x is a vector containing features that describe the word, for example, the parts of speech, tenses, among other lexeme string features described in Chapter 3, and features that describe the learner's actions on the app, namely, time since the word was last seen, history of how many times the word was last seen, and how many times the learner got the word correct.
2. This model focuses on x where solely the word features are assessed and their impact on the ability of recall. The word features are lexeme string features, word and lemma lengths.
3. Lastly, this model focuses on how exposure features on the app impact the ability of recall. The exposure features are delta, number of sessions the word was last seen and history of recalling the word correctly in those sessions.

4.2.1 Model 1

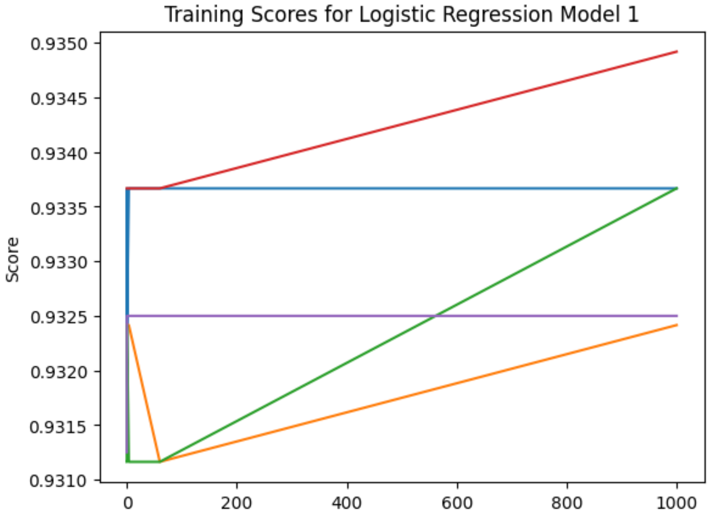


Figure 4.1: Training Scores for Model 1

Since the model is quite large, I focus on 1000 samples to calculate the training and validation scores. As can be seen from Figures 4.1 and 4.2, the training score tends to be higher than the validation score with mean training score being 0.9325 and mean validation score being 0.9318. This implies that the model does not pose issues of overfitting. This comparison of training and validation scores indicate that this model on its own seems to fit the data well. Next, let's hone in on the score which will be revisited at the end of this section. The score achieved with this model is 0.927.

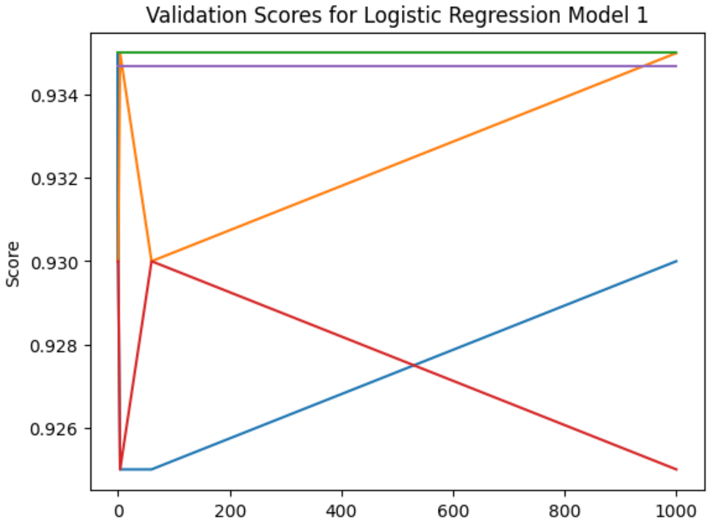


Figure 4.2: Validation Scores for Model 1

4.2.2 Model 2

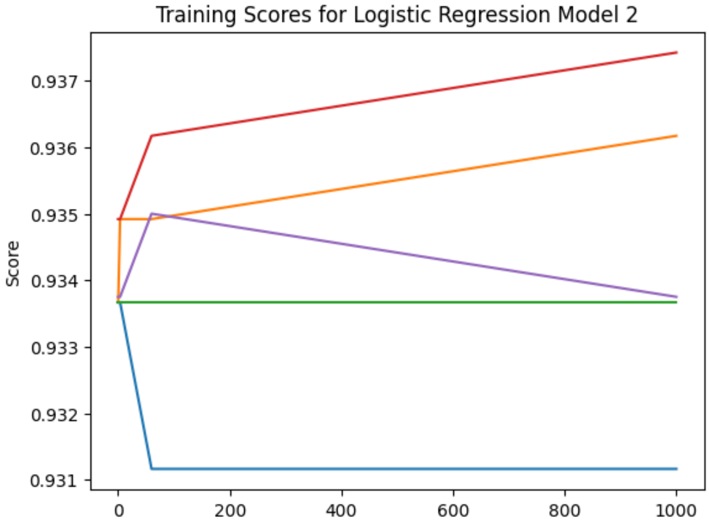


Figure 4.3: Training Scores for Model 2

Similar to Model 1, for convenience let's focus on 1000 samples to calculate training and validation scores. The mean training score is higher than the mean validation score, with the mean training score being 0.9340 and the mean validation score being 0.9330, with both scores being incredibly close to each other. Additionally, the model's score is 0.927, quite similar to Model 1. Here, this indicates that there is no overfitting as the model seems to fit the data well.

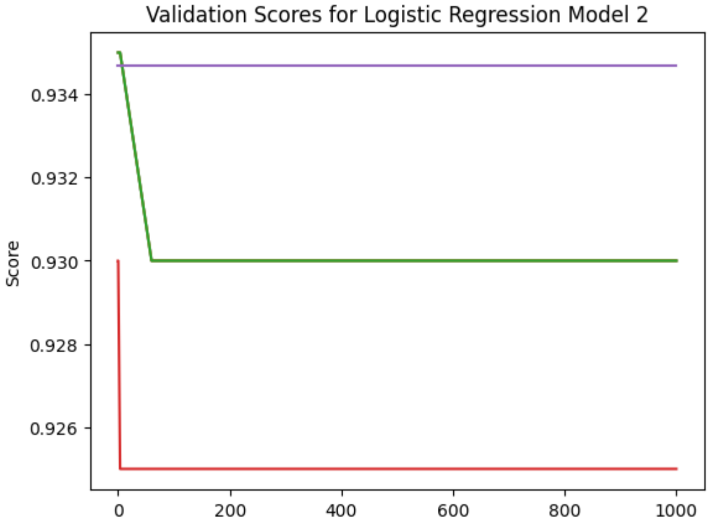


Figure 4.4: Validation Scores for Model 2

4.2.3 Model 3

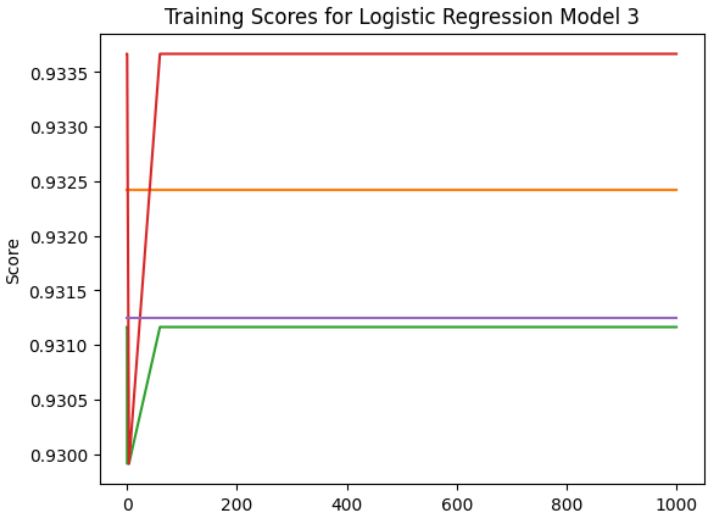


Figure 4.5: Training Scores for Model 2

The model's score of 0.927 is in line with the first two logistic regression models. Moreover, the mean training score, at 0.9321, is higher than the mean validation score, at 0.9317. Since both models scores are close to each other, there does not seem to be an issue of overfitting.

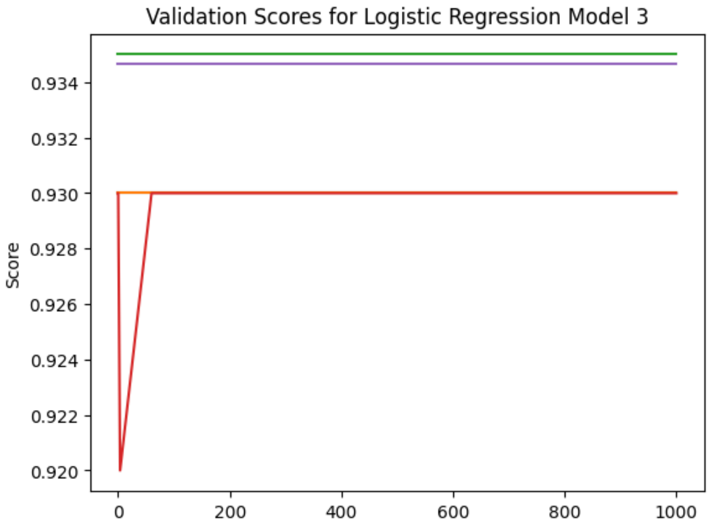


Figure 4.6: Validation Scores for Model 3

4.2.4 Results

Features	Coefficients
singular	0.56
determiner	0.41
word_length	0.36
verb	0.19
definite	0.18

Table 4.1: Top 5 Features by Magnitude

Since Models 2 and 3 have similar accuracy scores to Model 1, let's focus on Model 1 and assess its coefficients. The results in the above table are quite interesting. It can be seen that words that are in singular form, determiners, definite, or verbs increase the chances of recalling a word, similarly word length also increases the chances. Notably, delta and history of exposure variables do not have that large of an impact on the probability of recall according to this model.

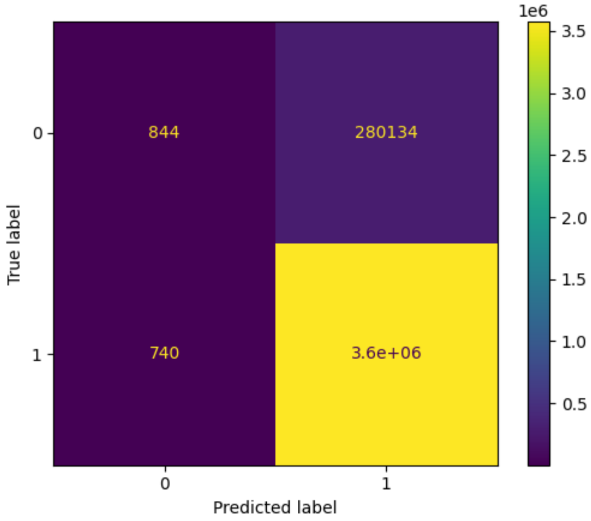


Figure 4.7: Confusion Matrix for Logistic Regression

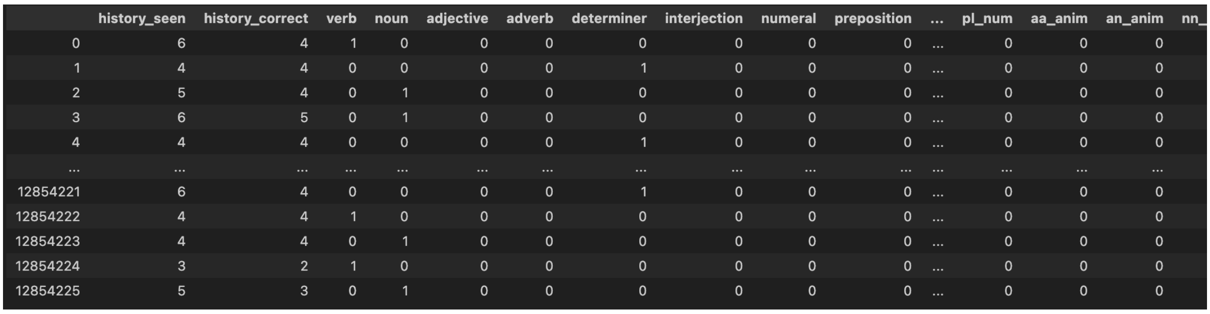
Additionally, the Logistic Regression model of all features tends to predict and recall quite well as can be seen from the resulting confusion matrix indicating that it is a strong contender for the best fit. The model is able to predict 1 or that the probability of recall is greater than or equal to 0.5, where the true label is also 1 for over 3 million observations. However, it can be seen that for observations where the true value is less 0.5, the model is predicting incorrectly for nearly 300,000 observations.

4.3 Half Life Model

4.3.1 Feature Engineering

As opposed to the logistic regression models in 4.1, I created an X value with the features: history_seen, history_correct, lexeme_string categorical variables that identify the part of speech, tense, and other modifiers, as well as word length and lemma length. Notably these X values do not contain delta and comprise of 65 features. Delta will be called in the model separately as a numpy array.

Learning Features Without Delta:



	history_seen	history_correct	verb	noun	adjective	adverb	determiner	interjection	numeral	preposition	...	pl_num	aa_anim	an_anim	nn_
0	6	4	1	0	0	0	0	0	0	0	...	0	0	0	
1	4	4	0	0	0	0	1	0	0	0	...	0	0	0	
2	5	4	0	1	0	0	0	0	0	0	...	0	0	0	
3	6	5	0	1	0	0	0	0	0	0	...	0	0	0	
4	4	4	0	0	0	0	1	0	0	0	...	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
12854221	6	4	0	0	0	0	1	0	0	0	...	0	0	0	
12854222	4	4	1	0	0	0	0	0	0	0	...	0	0	0	
12854223	4	4	0	1	0	0	0	0	0	0	...	0	0	0	
12854224	3	2	1	0	0	0	0	0	0	0	...	0	0	0	
12854225	5	3	0	1	0	0	0	0	0	0	...	0	0	0	

Figure 4.8: An Excerpt of X

I start off my data preprocessing by splitting the dataset into 80% of data that will be used as new data for the Random Forest Regressor and 20% will be used as data for the Half Life Model. While this is a large amount for the Random Forest Regressor I will explain this further in the next section.

4.3.2 Model

4.3.2.1 Random Forest Regression

Settles and Meeder in their 2016 paper detail out that when $p = 0.5$, which can also be written as $p = 2^{-1}$, delta equals half life. In this model I filter the new data for when the continuous probability of recall variable is equal to 0.5. Through this I am able to get the half life of each of the learning traces and explore the relationship between the learning features and half life using Random Forest Regression. I chose Random Forest Regression due to its ability to have lower variances (Scikit-learn, n.d.).

Unfortunately, the R squared score is only 0.01, I tried to increase the size of the dataset, however, of 80% of the dataset, there were only about 360,000 learning traces with probability of recall equal to 0.5. Regardless of the score, I will use this model to predict the half life values of the old data to be used in the Half Life Model.

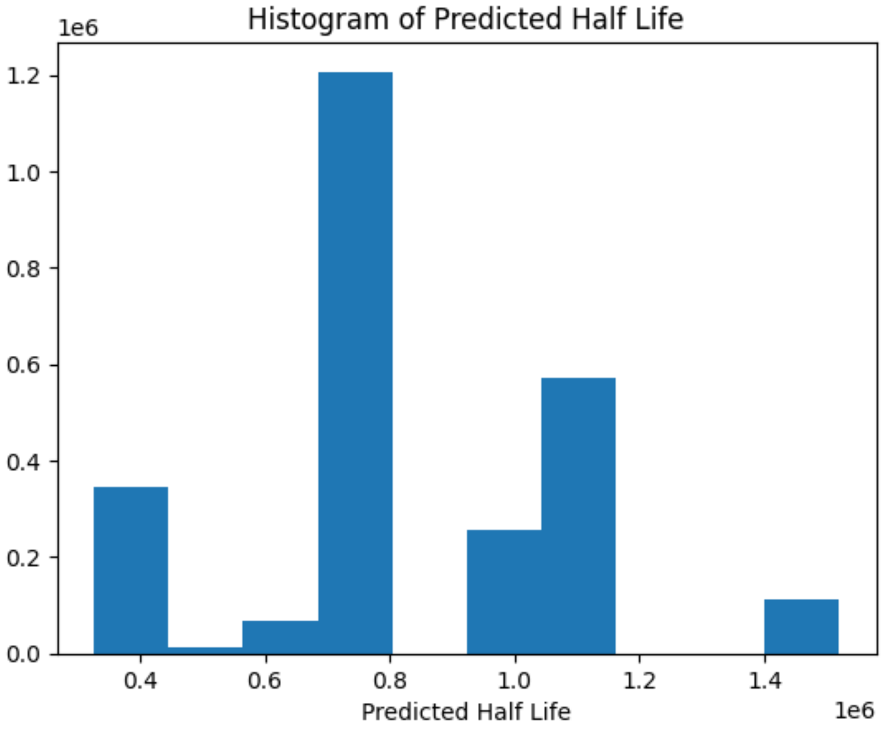


Figure 4.9: Histogram of Predicted Half Life

4.3.2.2 Half Life Model

Using sklearn's Base Estimator and Classifier Mixin, I was able to develop an sklearn classifier for a half life model (Scikit-learn, n.d.). Within the half life class, below are the two main functions:

1. Initialization function:

   It initializes values for delta.

2. Predict function:

   It utilizes Equation 2.1 to predict the probability of recall. By initializing an empty array, p, the function then runs a for loop for the 2.6 million instances in the old data set. It takes in the half life calculated using Random Forest Regression and then calculates the probability of recall at the ith instance using Equation 2.1, using delta at the ith instance and the half life.

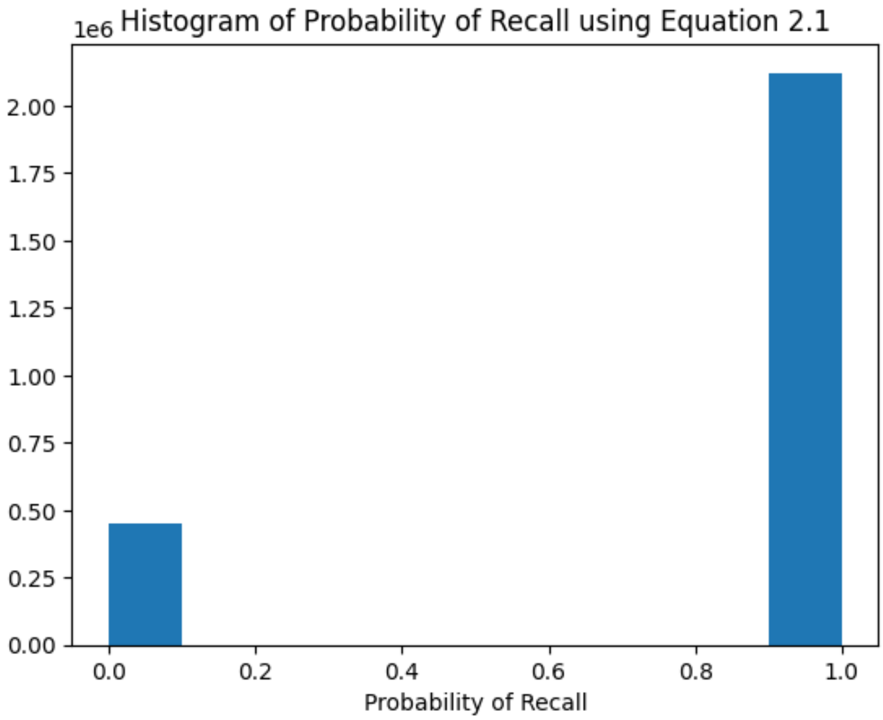


Figure 4.10: Histogram of New Probability of Recall

4.3.2.3 Results

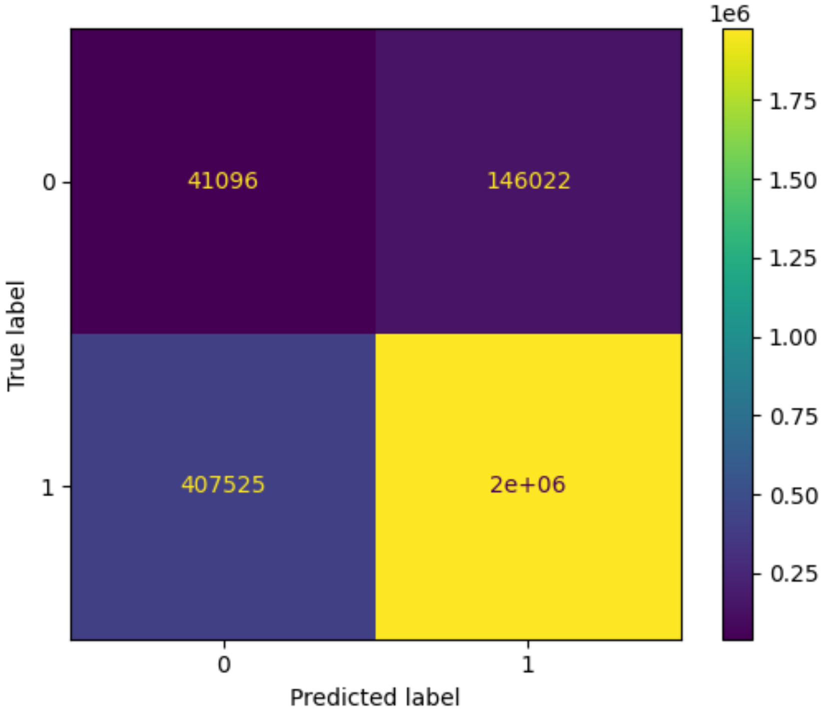


Figure 4.11: Confusion Matrix: Half Life Model

With regards to the Random Forest Regressor Model, only one of the features has a particularly high feature importance: history_seen. This suggests that history of exposure does play a large role in determining the half life of a word. This begs the question whether there are other exposure metrics that could be recorded in second language acquisition to understand how the half life of a word evolves. While, the first model could be improved, surprisingly, its half life predictions and subsequent computations of probability of recall indicate a high accuracy score. Its confusion matrix indicates that it is quite good at predicting whether the word would be recalled. However, it is important to note that, as seen in Figure 3.8, the dataset has more data points indicating recall, compared to not.

4.4 Other Models

4.4.1 Decision Tree Classifier

These classifiers can predict variables based on their decision-making (Scikit-learn, n.d.). I fit the decision tree model with all predictors including delta to predict the target categorical variable: probability of recall or ability to recall. In the table below, a learner's exposure to a word in sessions and how recently it was seen as important features in this model with delta outperforming all the other features. Word and lemma lengths are also important features in this model; this indicates that longer words may stay in memory longer.

Features	Importance
delta	0.67
history_seen	0.08
history_correct	0.07
word_length	0.05
lemma_length	0.04

Table 4.2: Top 5 Features by their Importance: Decision Trees

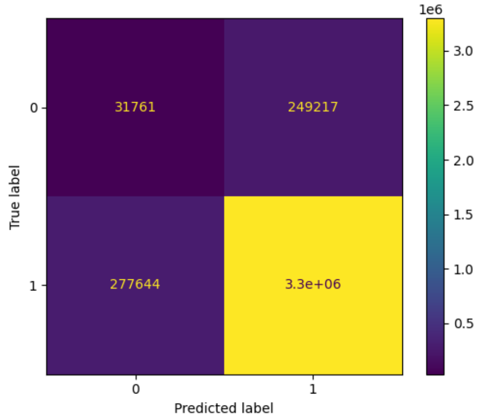


Figure 4.12: Confusion Matrix: Decision Trees

The model indicates good fit as can be seen in its ability to predict 3.3 million traces accurately. However, for nearly 300,000 traces where the probability of recall was greater than or equal to 0.5, the model predicted those inaccurately and a similar magnitude for probability of recall less than 0.5.

4.4.2 Random Forest Classifier

In random forests models, trees are developed through bootstrapping (Scikit-learn, n.d.). Similar, to the decision tree model, I fit the random forest with all predictors including delta. After running the model we see that delta, the number of times the word was seen previously, the word being a preposition, in infinitive tense, and the user guessing the word accurately in the past, were the most important features. These are in line with theory indicating that how often a word was seen and the ability to recall the word in prior sessions would contribute to the probability of recall. Moreover, the lexeme string features are words like have, among, near, which one would presume appear in various sentences and therefore, improve ability of recall. Additionally, the confusion matrix shows that there is room for the model's scores to be misleading as it is more likely to predict that the word is recalled.

Features	Importance
delta	0.29
history_seen	0.11
preposition	0.10
infinitive tense	0.09
history_correct	0.06

Table 4.3: Top 5 Features by their Importance: Random Forests

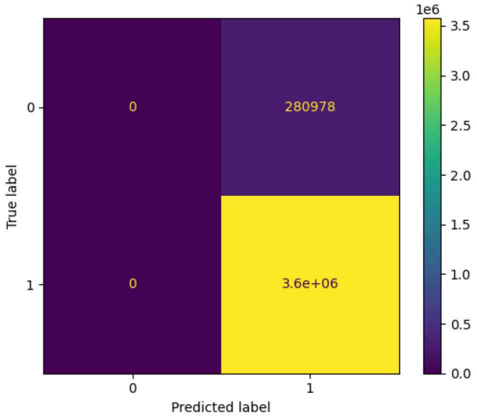


Figure 4.13: Confusion Matrix: Random Forest

CHAPTER 5

Model Comparisons & Discussion

	LR	**HL**	**Tree**	**Forest**
Accuracy Score	0.93	0.78	0.86	0.93
Precision Score	0.93	0.93	0.93	0.93
Recall Score	1.00	0.83	0.922	1.00
Mean Absolute Error	0.07	0.21	0.14	0.07

Table 5.1: Comparing the different Models

After running all four models, logistic regression (LR) and random forest (Forest) have the highest accuracy, precision, and recall scores, as well as lowest mean absolute error. Both these classifiers demonstrate through this that out of the four, they seem to be better candidates. The half life model (HL) lags behind in accuracy and recall scores, while having the largest mean absolute error. It does, however, have a high precision score. While the decision tree (Tree) model does have relatively better scores than the half life regression model, it lags behind logistic regression and random forest.

Interestingly, logistic regression focused more on word features to predict recalling a word, while random forest and decision tree had a mix of features. The random forest regressor, however, did list an exposure feature as an important feature. Moreover, the word features logistic regression and exposure features logistic regression models did not differ much in accuracy.

CHAPTER 6

Conclusions & Limitations

Using logistic regression to predict probability or ability of recall could improve second language acquisition. There are, however, limitations to this research. For example, the learning traces, while giving us valuable information about a student's exposure to a certain word, do not give much information regarding how that word was seen. Was the student prompted to match a word with an image or asked to decipher the word from an audio of an individual speaking the word? How the word was seen in historical sessions and the current session is integral to understanding how individuals learn languages.

Additionally, we do not know the motivations of the students using Duolingo. Are students using this to supplement their L2 education, as a game, or as the sole source of their L2 education?

Moreover, while the dataset has vital information regarding the language of the user interface it lacks student demographics. It may be the case that many students are using the app in English to learn other languages, while English is not their primary language. Does this, therefore, increase confusion when it comes to understanding the second language they are hoping to acquire? Other metrics that may be vital are students' age, how long they have used the app, what other languages they have learned, on the app and otherwise, whether they are multilingual or not.

Furthermore, this dataset does not contain languages like Hindi, Arabic, Korean, and Chinese, that have vastly different scripts than the languages currently in the dataset.

For students who are bilingual or multilingual, it may also be the case that some languages are easier to learn than others owing to having practiced learning languages before.

CHAPTER 7

Further Research Potential

As mentioned in the previous chapter, there are various features that would require additional research and may help better predict outcomes. Some suggestions may be incorporating existing models that have higher accuracy scores to improve overall accuracy scores. These models could be used to predict half life and subsequently, probability of recall. The idea of a half life of a word can also transfer to other elements of sight and sound. For example, the sound of a word could also elongate its half life. While metrics like seeing a word and word length were part of this thesis, there is potential of word associations increasing half life as well. For example, it may be the case an individual might be more likely to remember a red house and therefore, use that association to remember the word house.

Additionally, there is potential for more research on how half life could change with age, demographics and behaviors. In this thesis, I researched a few behavioral aspects, for example, exposure to the word and time since the word was last seen. However, there are other behavioral aspects that could be researched. Is there a particular kind of tense or part of speech that helps individuals retain a word in their memory? Do words sounding a certain way trigger a behavioral response and therefore, help in increasing the half life of a word? This research, while, having highlighted the importance of some predictors could benefit by incorporating more metrics.

CHAPTER 8

Appendix

Adjective	**Lexeme Code**	**Surface form / Lemma**	**Language**
Acronym	*acr*	tv / tv	English
Comparative	*comp*	worse / bad	English
Demonstrative	*dem*	those / that, this / this	English
Enclitic	*enc*	ce / ce, se / se	Italian
Interrogative	*itg*	which / which, who / who	English
Object	*obj*	me, them, her	English
Ordinal	*ord*	third / third, first / first	English
Possessive	*pos*	yours / yours, its / its	English
Proclitive	*pro*	si / si, la / lo	Italian
Quantifier	*qnt*	some / some, few / few	English
Reflexive	*ref*	ourselves / ourselves	English
Relative	*rel*	that / that, whose / whose	English
Synthetic	*sin*	green / green, white / white	English
Subject	*subj*	we, she	English
Superlative	*sup*	best / good, worst / bad	English
Tonic	*tn*	something / something	English

Cases	**Lexeme Code**	**Surface form / Lemma**	**Language**
Gentitive	*gen*	’s / ’s	English
Location	*loc*	England / America	English

Person Type	**Lexeme Code**	**Surface form / Lemma**	**Language**
First Person	*p1*	we, i, am	English
Second Person	*p2*	you	English
Third Person	*p3*	he, drinks / drink	English

Gender	**Lexeme Code**	**Surface form / Lemma**	**Language**
Feminine	*f*	she	English
Masculine	*m*	he, him	English
Masculine / Feminine	*mf*	we, yours	English
Gender Neutral	*nt*	it	English

Number	**Lexeme Code**	**Surface form / Lemma**	**Language**
Singular	*sg*	scale / scale, water / water	English
Singular / Plural	*sp*	its / its, yours / yours	English
Plural	*pl*	women / woman	English

Animacy	**Lexeme Code**	**Surface form / Lemma**	**Language**
Animate	*aa*	whose / whose	English
Animate / Inanimate	*an*	that / that	English
Inanimate	*nn*	what / what	English

Other Modifiers	**Lexeme Code**	**Surface form / Lemma**	**Language**
Anthroponym	*ant*	bois / bois	French
Apostrophe	*apos*	' / '	English
Definite	*def*	the / the	English
Indefinite	*ind*	an / a, a / a	English

References

Babbel. (n.d.). *It starts here.* https://www.babbel.com/about-us

Blanco, C. (2023). *2023 Duolingo Language Report.* Duolingo Blog. https://blog.duolingo.com/2023-duolingo-language-report/

Dabrowska, E. (2018). Experience, aptitude, and individual differences in linguistic attainment: A comparison of native and nonnative speakers. *Language Learning, 69*(S1), 72–100. https://doi.org/10.1111/lang.12323

Dabrowska, E., Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences, 28*(6), 604–615. https://doi.org/10.1016/j.langsci.2005.11.014

Duolingo. (n.d.). *About Us.* University Programs. https://university.duolingo.com/our-story-1

Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology* (H.A. Ruger C.E. Bussenius, Trans.). Teachers College, Columbia University. (Original work published 1885)

Hartshorne, J. K., Tenenbaum, J. B., Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Scikit-learn. (n.d.). *1.10. decision trees.* scikit. https://scikit-learn.org/stable/modules/tree.html

Scikit-learn. (n.d.). *1.1. Linear Models.* scikit. https://scikit-learn.org/stable/modules/linear_model.html

Scikit-learn. (n.d.). *1.11. ensembles: Gradient boosting, random forests, bagging, voting, stacking.* scikit. https://scikit-learn.org/stable/modules/ensemble.html

Scikit-learn. (n.d.). *Classifiermixin.* scikit. https://scikit-learn.org/stable/modules/generated/sklearn.base.ClassifierMixin.html

Seibert Hanson, A. E., Brown, C. M. (2019). Enhancing L2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning, 33*(1–2), 133–155. https://doi.org/10.1080/09588221.2018.1552975

Settles, B. (2016). *HALFLIFE-REGRESSION/LEXEME_REFERENCE.TXT*. GitHub. https://github.com/duolingo/halflife-regression/blob/master/lexeme_reference.txt

Settles, B. (2017, December 14). *Replication data for: A trainable spaced repetition model for language learning.* Harvard Dataverse. https://doi.org/10.7910/DVN/N8XJME

Settles, B., Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* https://doi.org/10.18653/v1/p16-1174

Zarrati, Z., Zohrabi, M., Abedini, H., Xodabande, I. (2024). Learning academic vocabulary with digital flashcards: Comparing the outcomes from computers and smartphones. *Social Sciences & Humanities Open, 9*, 100900. https://doi.org/10.1016/j.ssaho.2024.100900