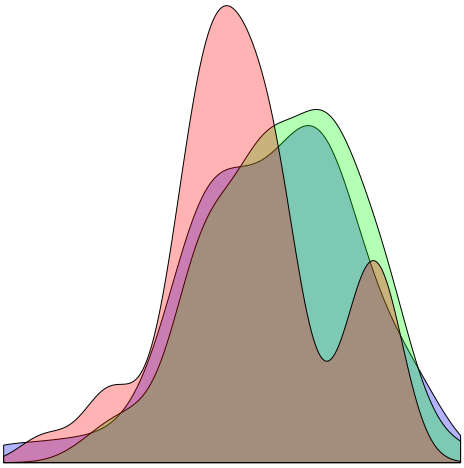
Utrecht University

Bachelor Thesis Artificial Intelligence

Predicting Dyslexia and Vocabulary Age



Anna Langedijk
5716128

7.5 ECTS

Supervisor: Dr. Hugo Schnack
Second reader: Dr. Tejaswini Deoskar

January 17, 2019

Abstract

Methods from the field of machine learning can be applied to automatically predict disorders such as dyslexia at an individual level. Early vocabulary is a plausible antecedent to later language difficulties. Previous research on this link has been mostly correlational and not predictive. In this study, productive and receptive vocabulary sizes measured in children between 17 and 35 months were used to predict their later dyslexia status. Linear support vector machines were trained to separate dyslexic subjects from nondyslexic subjects. Additionally, support vector regression was used to predict age from vocabulary. Dyslexia could not be reliably predicted: the maximum balanced accuracy was 58% for the group at 23 months old. The vocabulary age models did have a good fit (for the best model, $R^2 = 0.686$) and performed well on unseen data. Moreover, the age models predicted dyslexic subjects to be up to two months younger than their nondyslexic peers. This difference was however not enough to predict eventual dyslexia status. In conclusion, infant vocabulary was a weak predictor of dyslexia. Using data from multiple points in time might increase predictive performance, as vocabulary trajectory is nonlinear and differs in children with dyslexia. Similarly, the "vocabulary age gap" could be examined further, since vocabulary age models predicted dyslexics to be younger even without prior knowledge of dyslexia.

Contents

1 Introduction

1.1 Machine learning within psychology

Diagnosis of cognitive or psychological disorders is often time-consuming and requires extensive experience. Machine learning techniques can be used as methods to construct models which automatically give predictions when given relevant data of people with or without the disorder to be detected. Applying statistical methods like these can speed up the diagnostic process, which is especially desired in disorders when early intervention is beneficial. It can also help experts find anomalies and thus increase diagnostic accuracy. Models that learn from data alone, without human preprocessing, might give new insights that were not considered a priori.

Using statistical methods for these tasks is a relatively new and promising approach, but there are many things that can go wrong when applying machine learning in this domain of science (Kassraian-Fard et al., 2016). Data of any sort of medical significance is expensive and difficult to collect. As a result, this field suffers more than others from (very) small sample sizes. Models that are trained with limited data usually do not generalize well to unseen data. Statistical models often also require tuning parameters which are not calculated fairly because of limited available data (Cawley and L. C. Talbot, 2010).

A second problem is that of data imbalance. Often, when researching a specific disorder, there are fewer subjects to be found with that specific disorder in comparison to control subjects. This also makes it harder to assess classification models properly, as it is tempting for a statistical model to classify everyone as healthy just because that is the majority class. Additionally, when using a predictive model, how certain does it have to be of its prediction? Do we allow many false positives just to be certain we pick up on all cases, or is it better to be specific and allow more false negatives? These problems need to be considered when using predictive models in practice.

Dyslexia is a cognitive disorder characterized by difficulties in reading and other language abilities. It is classified as an early onset developmental disorder in the DSM-V. Early language skills such as vocabulary, therefore, can be viewed as a plausible antecedent to dyslexia (Thompson et al., 2015). However, collecting vocabulary data is time-consuming and may be inaccurate since it is often reported by parents. In addition to that, dyslexia is estimated to occur in about 10 to 13% of the population (Chen et al., 2017). There is bound to be some imbalance in the data. Using robust machine learning techniques, is it still possible to predict dyslexia on an individual level?

1.2 Overview

At the start of chapter 2, literature on the prediction of dyslexia and the link between early vocabulary and dyslexia is reviewed. In the later sections of this chapter, the theory behind the algorithms and evaluation metrics that will be used is discussed, keeping the challenges of this particular study in mind. Methods will be discussed in chapter 3, including a description of the data sample and its features. Lastly, results for all models will be reported (chapter 4), followed by a discussion of these results (chapter 5) and a conclusion (chapter 6).

2 Background

2.1 Literature

Infant vocabulary can be separated into two categories: expressive vocabulary (or productive vocabulary) is the set of words a child both produces and understands. A child's receptive vocabulary consists of all words a child understands but does not yet produce.

Although correlations between expressive and receptive vocabulary measured at 16 to 24 months and school-age reading performance were highly statistically significant, their predictive ability at an individual level was not. There seems to be a low developmental stability in early vocabulary knowledge (Duff et al., 2015b,a). Many children with very low expressive vocabularies caught up later, and many children with normal vocabularies did have reading difficulties later on. In other words, there was a high rate of false negatives and a high rate of false positives predicting later language difficulties from just vocabulary. Differences in expressive vocabulary measured at 18 months were statistically significant for later reading outcome, but by no means determinative. When accounting for other factors, such as gender and familial risk, the vocabulary assessment at such a young age might assume more importance.

Dyslexia has an early onset, but its effects start to be noticed only when the child is already enrolled in school. Reliable screening for risk of dyslexia in pre-school years could enable early intervention. Studies so far have rarely used pre-school language abilities to predict individual (risk of) dyslexia. Whereas the difference at group level between control and dyslexic children is reported, most did not look at predictions on an individual level.

Thompson et al. (2015) did predict individual risk of dyslexia given vocabulary and grammar knowledge in various pre-school age groups. The earliest models, at 42 months of age, performed poorly. When set at a cutoff that gave 90% sensitivity, the specificity was very low (30%). Language skill was not a significant predictor of dyslexia until the age of 5 years. Again, infant vocabulary did not seem to be stable across childhood: the closer to school entry, the better the predictive models were. Including familial risk of dyslexia yielded somewhat better results, but the study strongly suggests that screening for language problems at just 42 months provides little useful information with respect to later dyslexia.

Thompson et al. (2015) used logistic regression to predict individual risk, but the study was still largely correlational rather than predictive. They did not test cases outside of the sample that was used to build the model (Chen et al., 2017). A further limitation of this study is that the sample included only high-risk children, either children at familial risk for dyslexia or with a specific language impairment.

A study that did apply proper machine learning methods is Chen et al. (2017). Children's expressive and receptive vocabulary between the ages of 17 and 35 months was measured and used to build classifiers predicting familial risk of dyslexia[1]. Unlike Thompson et al. (2015), this study had access to a control group not specifically at risk for dyslexia. Model performance was assessed on subjects outside of the training sample. The best model performed with a 68% accuracy (and a 68% balanced accuracy) on the 19 to 20 month model. Although this study did use cross validation to report this accuracy, it still did not predict dyslexia itself, only familial risk, as the dyslexia status of the subjects was unknown at the time.

A subset of these children were followed up and eventual dyslexia status was assessed. This yielded

[1] Meaning at least one parent is reading impaired, and one first-degree family member is reading impaired.

three groups: TD (typically developing children, no dyslexia and no familial risk), FR-ND (children at familial risk without dyslexia) and FR-D (children at familial risk with dyslexia). There was no significant group difference between TD and FR-ND children. However, for both expressive and receptive vocabulary, the dyslexic children did consistently score lower than both nondyslexic groups. The effect sizes were all small to medium. Infant vocabulary, therefore, may function as an additional risk factor for the development of dyslexia, but it is still only weakly related (van Viersen et al., 2017).

This study also noted that although early vocabulary studies regarding dyslexia are longitudal, the focus is rarely on vocabulary trajectory. Vocabulary development is not entirely linear during infancy: a vocabulary spurt in the second year of life has been regularly observed (Hamilton et al., 2000). In FR-Dyslexic children, this vocabulary spurt had a later onset for expressive vocabulary. For receptive vocabulary, the spurt seemed to be characterized by a lower initial growth followed by a weaker deceleration at 29 to 35 months (van Viersen et al., 2017). Thompson et al. (2015) reported accuracy for models fitted for every seperate age group, only accounting for vocabulary size differences at specific points in time, and not for trajectory differences. Classifiers that have knowledge of multiple age groups might perform better.

Aims of this study

Machine learning methods have not been used to predict dyslexia from early vocabulary. Most studies focus on descriptive statistical methods, but so far, predictive models have been used to successfully predict other disorders (Kassraian-Fard et al., 2016). Prediction of familial risk at dyslexia yielded a maximum performance of 68% accuracy at 19 to 20 months (Chen et al., 2017). Since the effect size between dyslexics and nondyslexics is much larger than that between FR and typically developing (TD) (van Viersen et al., 2017), predicting dyslexia instead of FR at an individual level could yield similar or better results. The goal is to not just fit the best possible model, but use robust validation methods and consider the class imbalance when reporting on important metrics (Kassraian-Fard et al., 2016).

The second aim is to predict age at an individual level. How big is the gap between real age and predicted or "vocabulary" age in dyslexic children[2]? Likewise, is there a significant difference between the control, FR-ND and FR-D groups in vocabulary age? Although this model predicts age and not dyslexia, a persistent difference between real age and vocabulary age might give insight into the dyslexia status of the developing child.

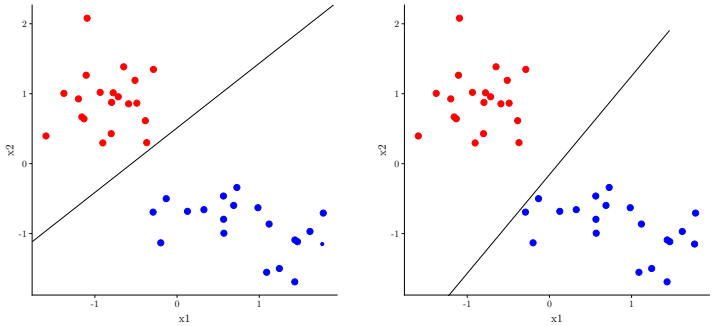
In the following sections, a theoretical overview of the used techniques is given.

[2]Similar to the real-age/brain-age gap for schizophrenia in (Schnack et al., 2016)

2.2 Support Vector Machines

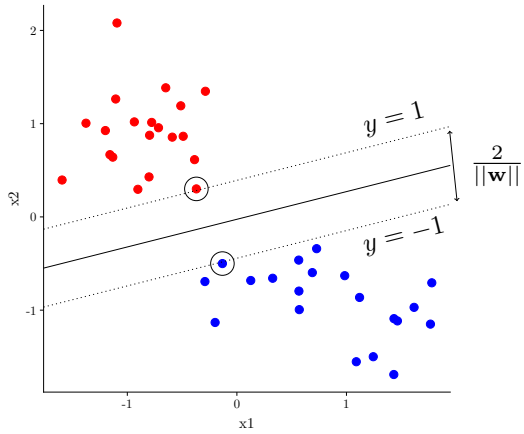
Support Vector Machines, first introduced by Cortes and Vapnik (1995), are a popular machine learning technique. SVMs are known to deal well with high dimensional and noisy data. It also deals well with correlated features. Vocabulary categories are of course highly correlated with each other, and these features can be noisy because of its manual collection via parental report. Additionally, in contrast to other ML techniques like neural networks, a linear SVM has fairly interpretable feature weights (Kassraian-Fard et al., 2016).

The goal of any classifier is to predict which of two or more classes some subject belongs to, given its features. A classifier has to separate the data into groups. The basic idea of a support vector machine (SVM) is that it is a maximal margin classifier. Consider the two-dimensional dataset in Figure 2.1: There are many ways it can be separated with a line, but not all of them would be considered good separators.



(a) A separating line

(b) Another separating line



(c) The optimal separating line

Figure 2.1: A linearly separable dataset in two dimensions. All of the three lines separate the data perfectly, but (c) does so while maintaining the maximum margin. The circled data points are the so called support vectors.

The goal of an SVM is to separate the data, but specifically to do so while maximizing the margin, the distance from the separating line to the nearest datapoint. In Figure 2.1a, for instance, the distance to the nearest red datapoint is very small. To find the optimal separating line (or optimal separating hyperplane in higher dimensions) we need an optimal weight for every dimension, plus the intercept b. Let $\mathbf{w}$ be our weight vector. For some datapoint $\mathbf{x}$ the output of our SVM is: $sign(\mathbf{w}^T\mathbf{x} + b)$. Suppose y is the actual label (equal to -1 or +1) for some $\mathbf{x}$. Notice that when our

classifier classifies $\mathbf{x}$ correctly, $y(\mathbf{w}^T\mathbf{x} + b) > 0$.

In order to be true to our first constraint (classify everything correctly), the following must hold for every data point x with label y: $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$. Secondly, the margin with width $\frac{1}{||\mathbf{w}||}$ must be maximized. This is equivalent to minimizing the inverse of the margin width. Our problem becomes:

$$\text{Minimize } ||\mathbf{w}|| = \frac{1}{2}\mathbf{w^T w}$$
$$\text{subject to } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

This can be reduced to a quadratic programming problem and solved accordingly.

2.2.1 Soft-Margin SVM

The aforementioned SVM works only on perfectly linearly separable data. When the data is not separable by a line, the constraint "classify everything correctly" cannot be resolved. A more general version of an SVM specifies that points may be classified incorrectly to a certain extent. This is realistic, because many datasets are not linearly separable, but a big margin is still preferred.
The extent to which data may be classified incorrectly can be described by the hinge loss function, which penalizes every data point according to how far away it is from the margin. Points are allowed inside the margin, but they are given a penalty. This way, correctly classified data points outside of the margin do not contribute to the error.
The amount of loss allowed is controlled by a new parameter C, also known as the cost parameter. Let ξ_n be the hinge loss at any point n. Our optimization problem now becomes:

$$\text{Minimize } \frac{1}{2}\mathbf{w^T w} + C \cdot \sum_{n=0}^{N} \xi_n$$
$$\text{subject to } y_n(\mathbf{w^T w x_n} + b) \geq 1 - \xi_n,$$
$$\xi_n \geq 0$$

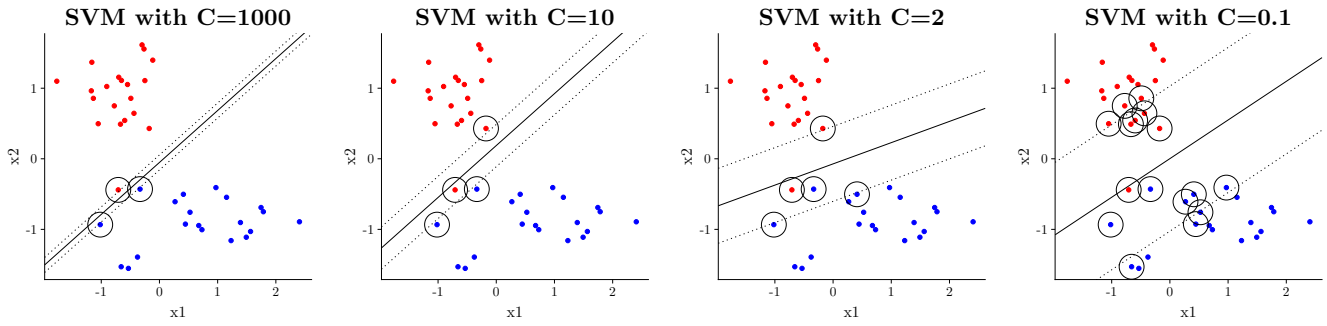


Figure 2.2: **Four soft-margin support vector machines**
All are trained on the same dataset, but with different values of C. Note that this dataset is linearly seperable, but the lowest red dot seems to be an outlier we want to ignore. The dotted lines next to the seperating line show the margin of the optimal separating hyperplane and a circled datapoint indicates that this point is a support vector. With $C = 1000$, the classifier is essentially a hard-margin SVM.

Notice that C controls the trade-off between minimizing the margin, and minimizing the error. When C is small, the second term is less important. The margin will be bigger, but the error higher. A very high C results in a small error, but at the cost of a small margin. When C is infinitely big,

the optimal hyperplane is the same solution as a hard-margin SVM would give us. In Figure 2.2 the influence of C can be clearly seen.
The cost parameter for SVMs has to be chosen empirically, usually from a range of values. In Figure 2.2, the optimal C seems to be about 2, but it is not hard to imagine that C has to be bigger or smaller for different datasets and -distributions.

2.2.2 Support Vector Regression

Typically, SVMs are used as classifiers. They can however also be used as regression models (Drucker et al., 1997). The classifier produced by an SVM, as described in the previous section, depends only on a subset of the training data, because the cost function disregards any points outside of the margin. For support vector regression, the same is true. We cannot ask for zero error (like we could with classification), since our output is continuous. What we can ask for is an error within an acceptable range, say an error per data point smaller than some $\epsilon > 0$. This "hard" version of SVR is formulated as follows:

$$\begin{aligned} \text{Minimize } & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to } & y - \mathbf{wx} - b \leq \epsilon, \\ & \mathbf{wx} + b - y \leq \epsilon, \\ & \epsilon \geq 0 \end{aligned}$$

Of course we cannot guarantee that the data is "linear enough" and the errors all lie inside the ϵ-range. Again, a cost parameter C is introduced, along with the hinge loss per datapoint ξ to allow for errors. This hinge loss is zero for every error smaller or equal to ϵ, and just like the classification hinge loss, increases linearly when it is further away from epsilon ("outside the margin").

$$\begin{aligned} \text{Minimize } & \frac{1}{2}||\mathbf{w}||^2 + C\frac{1}{N}\sum_{i=0}^{N}(\xi_i + \xi_i^*) \\ \text{subject to } & y - \mathbf{wx} - b \leq \epsilon + \xi_i, \\ & \mathbf{wx} + b - y \leq \epsilon + \xi_i^*, \\ & \xi^{(*)} \geq 0, \\ & \epsilon \geq 0 \end{aligned}$$

Now, there are two hyperparameters to optimize: C and ϵ. The "margin" of the SVM is now defined by ϵ, which creates a tube of width 2ϵ around the regression line, and any points classified within that tube do not contribute to the error. Since there is no real margin, C is now just a reguralization constant, with a lower C accounting for a flatter, more regularized model (Smola and Schölkopf, 2004). The trade-off between the two parameters is shown in Figure 2.3.

Since the outcome of our model is highly dependent on ϵ, it would be nicer to tune this parameter automatically. The most common variant of SVR used, ν-regression, does just this. When given a $0 \leq \nu \leq 1$, optimize ϵ as follows:

$$\begin{aligned} \text{Minimize } & \frac{1}{2}||\mathbf{w}||^2 + C(\frac{1}{N}\sum_{i=0}^{N}(\xi_i + \xi_i^*) + \nu\epsilon) \\ \text{subject to } & y - \mathbf{wx} - b \leq \epsilon + \xi_i, \\ & \mathbf{wx} + b - y \leq \epsilon + \xi_i^*, \\ & \xi^{(*)} \geq 0, \\ & \epsilon \geq 0 \end{aligned}$$

Another property of ν is that it is an upper bound on the fraction of errors, and a lower bound on the fraction of support vectors (Smola and Schölkopf, 2004). Often, as a default, $\nu = 0.5$ is chosen when there is no room for additional model tuning.

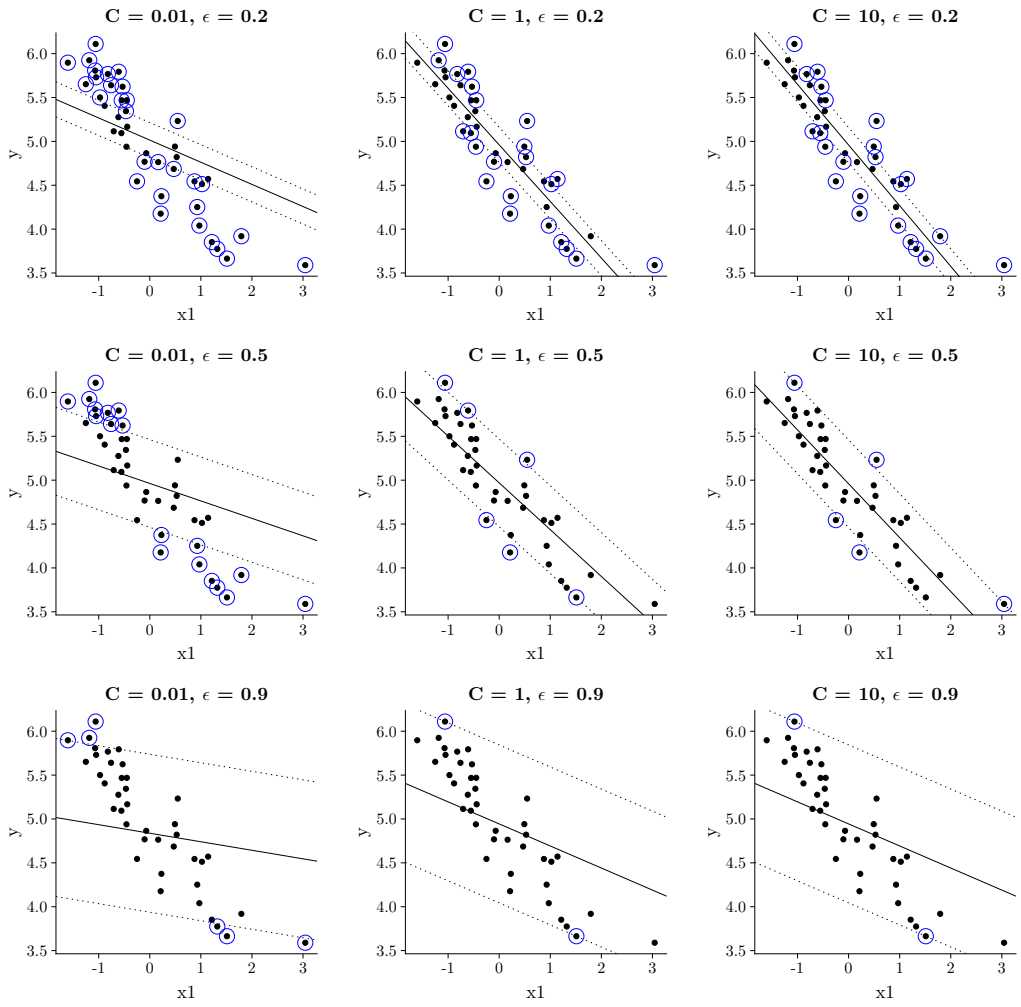


Figure 2.3: ϵ-**Regression for varying hyperparameters**
For some fixed ϵ, we allow ϵ error for any point. Nine regression models trained on the same (simulated) data for different values of ϵ, C. The support vectors (circled in blue) are the ones for which our model prediction has an error $> \epsilon$. Note that C determines the flatness of the model. Maximizing C seems like the best thing to do in this situation, but for nonlinear problems, optimizing C is less trivial.

2.3 Cross Validation

When assessing the performance of any model, it is not fair to do so on the same data that was used to train this model in the first place. The model might overfit on the data, meaning it does not generalize well on unseen data. A seperate dataset which is not used during training, called a test set, can be used to report on the performance in a fair way. Usually, the provided data is split up into a test set and a training set prior to any training. Data is trained on the training set and its performance evaluated on the separate test set. But with a limited amount of data, this process can be challenging. The larger the test set, the more accurate the performance assessment is for unseen data. But a larger test set also results in a smaller training set, which in turn makes this performance worse. When there is little data available, there are other techniques to evaluate model performance.
Another common way to assess performance is K-fold cross validation. The available data is split

up into K subsets of roughly equal size, called folds. For every fold, a model is trained on all other folds and then tested on the left-out fold. This way, all models are evaluated on data that they have not seen before, and we get to keep all of our data. When K is equal to the number of datapoints in a set, it is also known as leave-one-out cross validation (LOOCV). The model is then trained on all subjects except one, and evaluated on the left-out subject. Although LOOCV tends to have higher variance than regular K-fold CV, bias does decrease[3].

When training a soft-margin linear SVM, an optimal C-parameter has to be selected empirically. Choosing C based on the same data (i.e., the data the training set for a given fold) can easily lead to overfitting (Kassraian-Fard et al., 2016). Therefore, the model for that fold should be tuned first, with a nested cross validation procedure. This method is time-consuming but more robust than regular cross validation (Cawley and L. C. Talbot, 2010).

Nested Cross Validation:

1. Split data up into K outer folds
2. For every outer fold:
   (a) Create a train-data set of everything but the left-out outer-fold
   (b) Repeat R times: Split train-data up into K' inner folds and for every fold:
      i. Create a inner training set of everything but the left-out inner fold
      ii. Pick a C from list of available C's.
      iii. Train an SVM with this C on the inner train set.
      iv. Assess performance on the left-out inner fold (with a given metric, such as AUC or Accuracy)
      v. Average performance over all different C's
   (c) Evaluate which was the best C
   (d) Train a final model for with this C-value on the outer-fold train data
   (e) Evaluate the performance of this model on the left out outer fold
3. The estimated performance of an SVM on our dataset is the performance on all left-out folds.

This process is shown schematically in Figure 2.4.

[3]A high bias means the best possible model is still far from the golden standard. A high variance means models trained on different data (for instance, different folds) produce wildly different performances. Both bias and variance should be minimized, but there is a trade-off between the two.

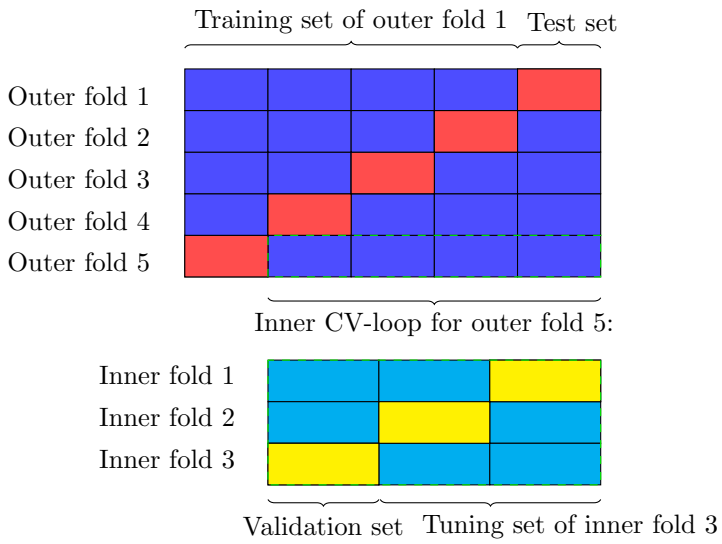


Figure 2.4: Nested Cross Validation with $K = 5$ and $K' = 3$. The performance of the crossvalidated model is measured using only the left-out outer folds, coloured in red.

2.4 Dealing with imbalanced data

Performance Metrics

The most well-known way to report on model performance is by measuring its accuracy, the amount of correctly classified data divided by the total amount of data.
However, when the data distribution contains a lot of points labeled positive, and almost none labeled negative, the positives will contribute a lot to the final accuracy. Even when every negative is incorrectly classified by our model, as long as the positives are all well, the accuracy stays high. The accuracy metric is thus misleading when dealing with imbalanced data (Kassraian-Fard et al., 2016; He and Garcia, 2009).

For binary classification, there are four ways a model can classify a datapoint. When a positive label means "this person has dyslexia" and a negative one means "this person does not have dyslexia":

- ⋆ TP or True Positive is when a dyslexic subject is correctly classified as being dyslexic.
- ⋆ FP or False Positive is when a non-dyslexic subject is incorrectly classified as being dyslexic.
- ⋆ TN or True Negative is when a non-dyslexic subject is correctly classified.
- ⋆ FN or False Negative is when a dyslexic subject is incorrectly classified as being non-dyslexic.

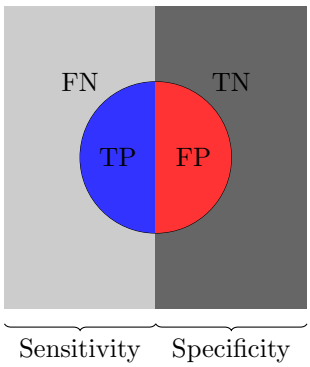


Figure 2.5: Performance Metrics

For many medical classification tasks, the positive group (in our case, dyslexics) is often (very) small and easy to ignore. But this is the very thing we are predicting, so it is important to pick up on those positive cases, although there are not many. The sensitivity metric, also known as recall or true positive rate, tells us how many of the positive class cases the model picked up on. With #TP standing for the total number of true positives, sensitivity

is defined as:

$$\frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}$$

When our classifier ignores positive subjects, this results in a high FN-rate, and in a low sensitivity.

With any medical classification, it is also important to avoid false positives(Thompson et al., 2015). This means we have to have a high specificity, also known as true negative rate, defined as

$$\frac{\#\text{TN}}{\#\text{TN} + \#\text{FP}}$$

When the classifier predicts many healthy subjects as dyslexic, "to be sure", the amount of false positives will be high and consequently, the specificity will be low.

Overall, there is a trade-off between sensitivity and specificity. This trade-off can be understood by looking at the ROC curve of a predictive model. The ROC curve shows the predictive power of a model at different thresholds. For an SVM, instead of using the sign of the prediction $sign(\mathbf{w}^T\mathbf{x}+b)$, the actual prediction probability (a number between 0 and 1) can be used. Ideally, all positive instances should have probabilities higher than every negative instance. A perfect classifier like this has some optimal threshold with 100% specificity and 100% sensitivity. Two ROC curves are shown in Figure 2.6. The area under this curve is a measure to assess the goodness of an ROC curve: how much predictive power does a model have, regardless of the final chosen threshold? For a perfect classifier the area under the ROC curve (AUC) is 1, as shown in the picture[4]. An AUC of 0.5 means no predictive power.

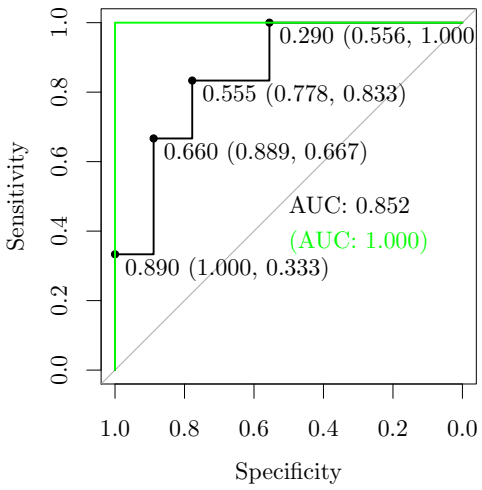


Figure 2.6: Example of an ROC curve for a small dataset (black), and perfect ROC curve (green), along with corresponding AUC measures. Some interesting thresholds and their coordinates in the specificity/sensitivity space are shown.

In a good model, both specificity and sensitivity have to be at least higher than 50%. The balanced accuracy metric is the mean of these values. Both metrics are just as important in the balanced accuracy, whereas with the regular accuracy metric they are important only proportionally to the balance of data. It is relevant to report on such measures and look at trade-offs instead of just accuracy (Schnack and Kahn, 2016).

[4]AUC can also be seen as the probability that some random positive datapoint has a higher probability than some random negative datapoint.

Sampling Methods

Reporting the results of a classifier model in a fair way is important, but during the training phase we can also account for the imbalance in our data.
Instead of using our imbalanced data set, we could sample the data and pretend that it is balanced. For big datasets, the majority class could be undersampled. For small datasets, we can oversample the minority class. This means that during training, the algorithm randomly samples the smaller class to be the same size as the majority class. Both classes are now equally important, but this method can lead to overfitting (He and Garcia, 2009), because it reuses datapoints during training.

Adding Class weights

Another way to account for class imbalance during the training phase is adding class weights. This means that during training, a heavier penalty is given for misclassifying the smaller group. For an SVM specifically, this means the C can be translated into two different costs, C_1 and C_2:

$$\text{Minimize } \frac{1}{2}\mathbf{w^T w} + C_1 \cdot \sum_{y_n=-1}^{N} \xi_n + C_2 \cdot \sum_{y_n=+1}^{N} \xi_n$$

There is no standard calculation of class weights. When the positive class is smaller, naturally, $C_2 > C_1$ should be true for optimal results. If the two classes are equally important, these weights can be chosen inversely proportional to the class distribution.

3 Methods

3.1 Sample Description

The sample consists of 212 Dutch children in total, for whom both familial risk status (FR) and dyslexia status is known. Their vocabulary was assessed at one or more stages, at 17, 23, 29 or 35 months of age. There were 147 children with data at all possible time steps.

Since more children from the at-risk group were followed up, there exists a class imbalance between FR and typically developing (TD) children. Moreover, only a small subset of the sample developed dyslexia, so there is an even higher imbalance between dyslexic and nondyslexic subjects. The balance of the data can be seen in Table 3.1. Notice that dyslexia is a subset of FR, no TD children developed the disorder.

	At-Risk	TD
No Dyslexia	92	69
Dyslexia	51	0

Table 3.1: Balance of all data

	At-Risk	TD
No Dyslexia	76	63
Dyslexia	40	0

(a) 17 months

	At-Risk	TD
No Dyslexia	81	63
Dyslexia	44	0

(b) 23 months

	At-Risk	TD
No Dyslexia	84	65
Dyslexia	48	0

(c) 29 months

	At-Risk	TD
No Dyslexia	87	59
Dyslexia	46	0

(d) 35 months

Table 3.2: Balance of data per age group

The children had their vocabularies assessed for words in 22 semantic word categories provided by the N-CDI (Zink and Dejaegere, 2002), developed for children between 8 and 30 months. Categories include "verbs", "connecting words", "places outside the house" and "animal names". A full list of categories can be found in the Results section in Figure 4.7. For each word in each category, parents reported exactly one of three options: "does not understand", "understands but does not produce" or "understands and produces". The third category will be referred to as productive vocabulary. The receptive vocabulary is the sum of the second and third category, yielding the total amount of words understood by the child. The average vocabulary trajectories are shown in Figure 3.1. For a detailed description and statistical analysis of this data, see van Viersen et al. (2017).

Additionally, data from (Chen et al., 2017) was used for some validation. This includes subjects from every age group for which only FR status is known, and two extra age groups of 18 and 19 to 20 months[1]. For the other ages, there is some overlap with the subjects from the main sample, but the data includes more TD children and was balanced for the study in question.

[1]The subsets of children of 19 and 20 months were small: classification models were trained for these two ages combined.

Age	17	18	19-20	23	29	35
TD	100	57	60	102	95	84
At-Risk	99	57	60	103	94	82

Table 3.3: Balance of data for subjects used in (Chen et al., 2017), including the additional age groups at 18 to 20 months. Their final dyslexia status is unknown.

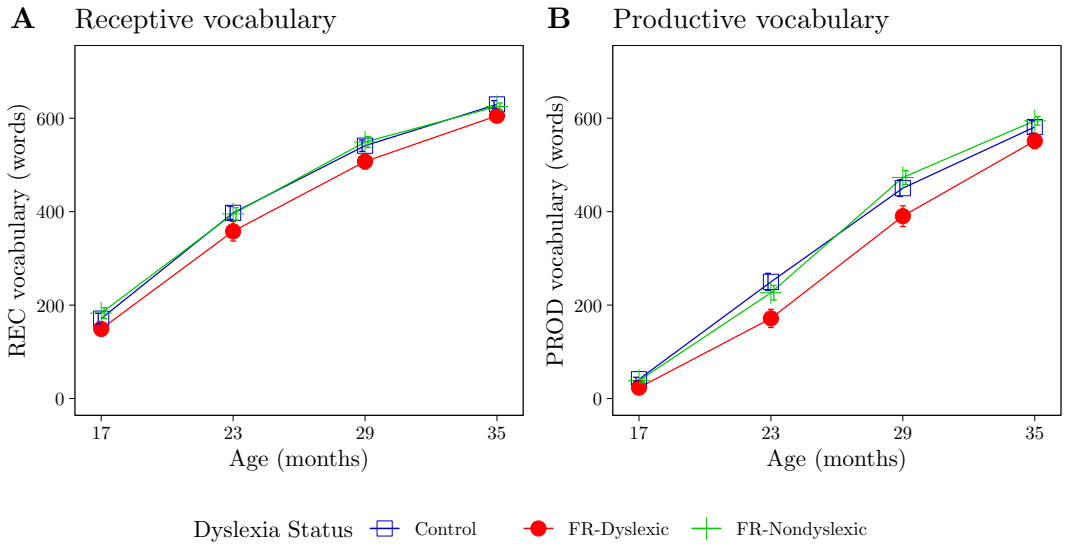


Figure 3.1: Mean and standard error of (A) total receptive vocabulary and (B) productive vocabulary at the four different timesteps. Dyslexics are constantly behind at the group level.

3.2 Predicting Dyslexia

For every age group, a soft-margin linear SVM was trained and evaluated using the data of only that age group. The classifiers separated either FR from TD, or dyslexics from nondyslexics[2]. The classifiers were trained on either receptive vocabulary (REC) or the productive vocabulary (PROD). This means 22 features per model. To increase the computational performance of the SVM, the data was scaled between 0 and 1 separately for every feature. This was done by first subtracting min(featureValue) and then dividing by max(featureValue) $-$ min(featureValue).

For evaluation, nested cross validation was used with $K = 5$ and $K' = 5$. The amount of inner repeats was 10. In the inner folds, C-values ranging from 0.001 to 10 were tuned ($C \in \{0.001, 0.05, 0.1, 0.25, 0.5, 1, 2, 5, 10\}$). The metric used for tuning was the area under the ROC.
$K = 5$ was chosen over a more traditional $K = 10$ because the dataset is imbalanced and small: when $K = 10$, every outer fold consists of around 18 subjects, of which four would be expected to have dyslexia. This can result in a very high variance of the sensitivity metric and overly optimistic AUC values[3].

To account for class imbalance, two different techniques explained in the previous chapter were used. Half of the SVM's were trained with class weights chosen inversely proportional to the data distribution, with the positive weight (either FR or Dyslexia) equal to 1 and the control weight equal to

$$\frac{\#\text{Positive-subjects}}{\#\text{Control-subjects}}$$

[2]FR-predictors were built in order to compare performance with the models built in (Chen et al., 2017).
[3]See also Figure 5.1 in the discussion section.

where "positive" is either FR or Dyslexia. The other half of the SVMs used oversampling for the tuning sets in every inner fold. Furthermore, the main metric to be reported is balanced accuracy instead of accuracy.

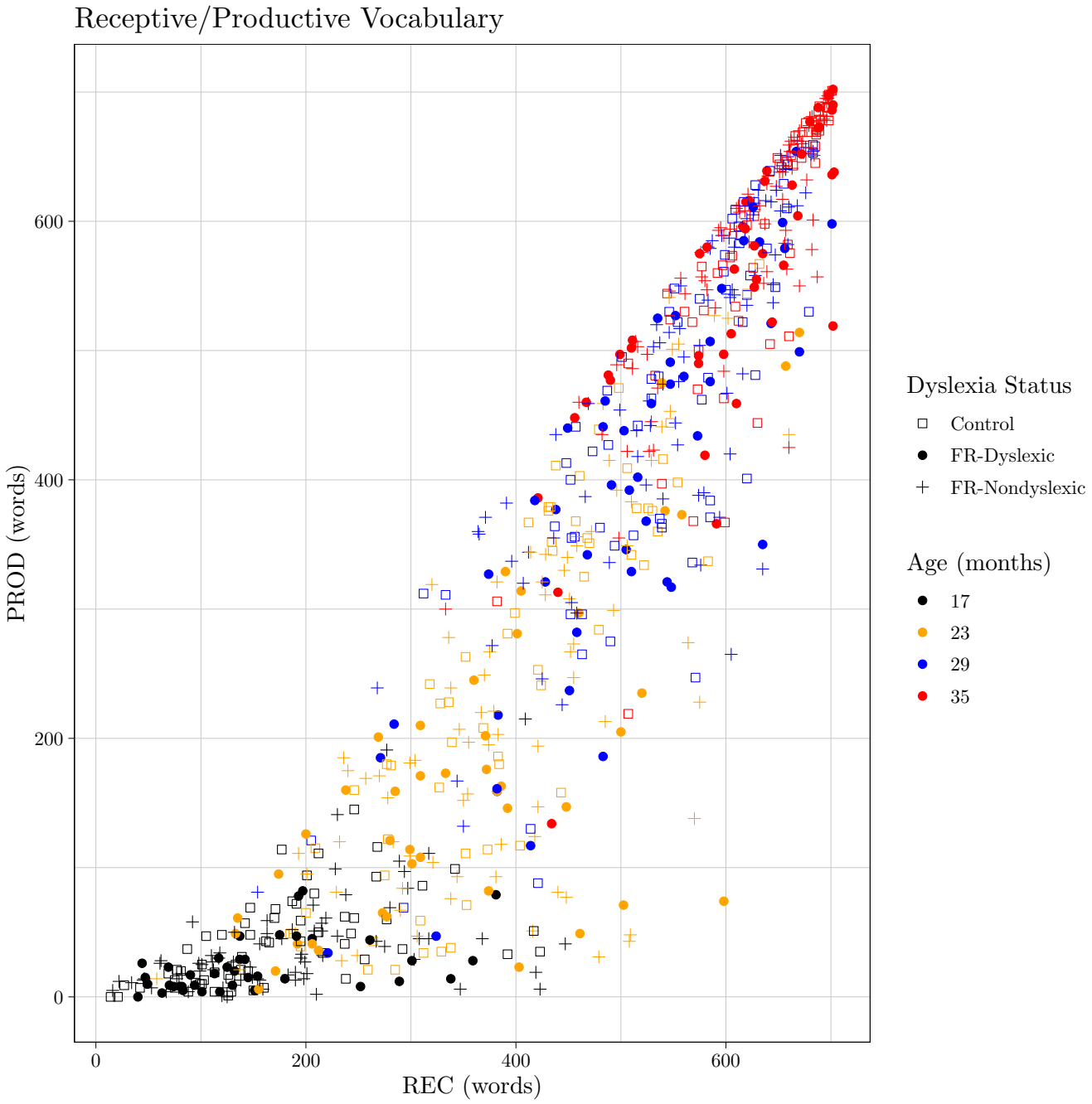


Figure 3.2: **Distribution of all data**
All datapoints plotted in "vocabulary space": receptive vocabulary on the x-axis, and productive vocabulary on the y-axis. Color signifies age, shape signifies dyslexia status. Notice that the diagonal cut-off means children do not produce any more words than they understand. It is crowded at the 35 month maximum, since the N-CDI is standardized for children up to 30 months, and many 35-month olds know almost all possible words.

3.3 Predicting Age

In order to predict age from vocabulary, ν-support vector regression with a linear kernel was used for different values of C ($C \in \{0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5\}$). Since the training data consists of at most 69 subjects, a fixed ν of 0.5 was chosen, which is the default in the e1071 package (Meyer et al., 2018). The C-parameter was tuned using LOOCV per subject. For every control subject, an SVR model was trained on all data of the other 68 control subjects and validated on the last subject. This means at most four datapoints for validation per subject. The goodness of fit (R^2) for this first model was determined using only left-out datapoints. Then, the C with the lowest RMSE was picked and used to train a final model on all control subjects. After that, the model was evaluated on the FR-children, and on the extra data from the 18 to 20 month old group[4]. Then, it was applied on the dyslexic subjects.
The vocabulary age models were trained on either REC or PROD features, and each feature was scaled between -0.5 and 0.5, first dividing by the maximum value of a feature, then subtracting 0.5.

Two models were trained with additional nonlinear features, where there were 44 features in total: for every feature x, x^3 was included as a separate feature. This was to account for the nonlinearity the vocabulary trajectory seems to have at this age (van Viersen et al., 2017). This nonlinearity may also be seen in Figure 3.1.

Additionally, all models were trained on just the real data, or the real data plus some generated data. This was in order to fight the effect of regression to the mean. Since the N-CDI measures vocabulary starting at 8 months, the age models, which have seen data starting at 17 months, have no knowledge of what happens before this age: children do not know any words yet. Two datasets of 60 control subjects were added on both sides, at ages 8 and 44 months. Data was generated separately for every vocabulary category, according to a normal distribution with $\mu = 0$ words and $\sigma = 0.2$ words[5]. The other dataset was added in the same manner at 35+9=44 months, with μ equal to the maximum amount of words in a category and $\sigma = 0.2$. By adding this data, the full vocabulary spurt might be seen more clearly. The average trajectory with generated data included can be seen in Figure 3.3. With generated data, the sigmoidal shape of the curve is accentuated.

Measuring vocabulary age difference

Dyslexic children are expected to have a lower predicted vocabulary age than their nondyslexic peers. To measure this effect size, FR-ND children were compared to FR-Dyslexic children. Including control subjects would mean including the training data, and a bigger effect size could also be explained by a worse generalization of the model and not the actual group difference. Effect sizes between FR-ND and FR-D were calculated using an unequal variances t-test (Welch's t and its corresponding p-value were reported). Additionally, the Hedges' g and mean difference in months was reported.

The regression models use subjects of all ages during training, and therefore contain some information about vocabulary trajectory. Seeing dyslexic children showed a low initial growth and more change in growth over time (van Viersen et al., 2017), a model that has knowledge of multiple ages might give better results.

To look at the predictive power of the vocabulary age model, an ROC curve was built for every age group with possible thresholds starting at the minimum predicted vocabulary age to the maximum predicted vocabulary age.

[4]Note that dyslexia status of these children is not known. With validation as a goal, these children were assumed to be nondyslexic.

[5]When a negative score was generated for a category, the absolute value of this was taken instead. For the maximum word groups, the opposite was done for the positive scores.

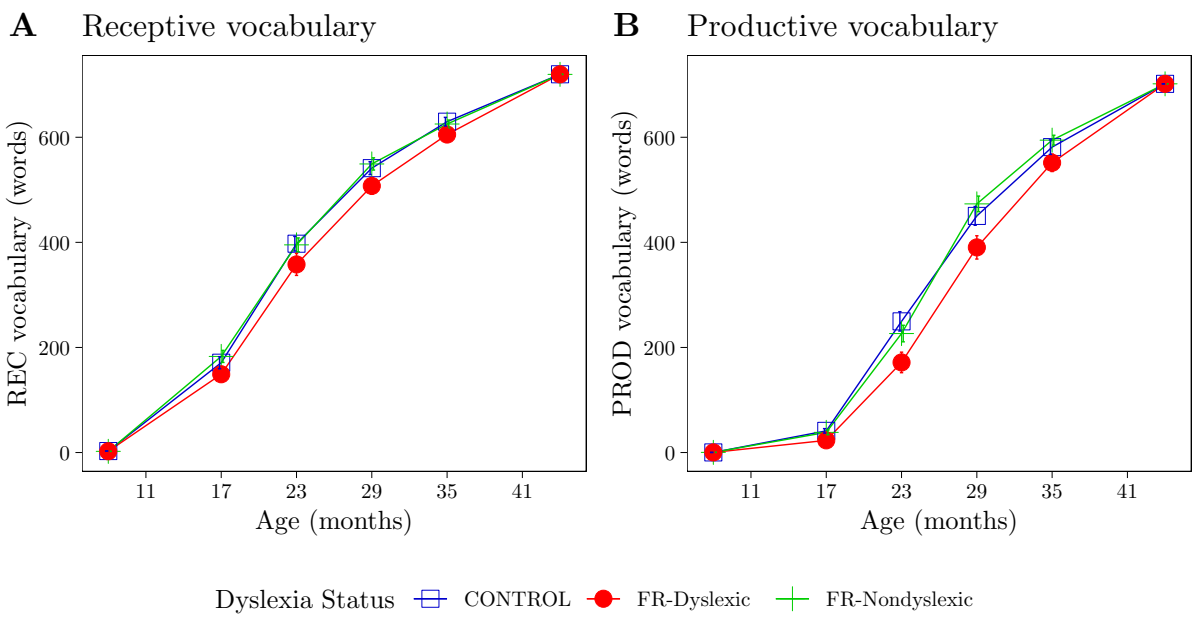


Figure 3.3: Mean and standard error of (A) total receptive vocabulary size and (B) productive vocabulary size, including the two generated age groups at 8 and 44 months. The youngest group is expected to know zero words, the eldest group is expected to know all possible words in the N-CDI. The curve of the vocabulary trajectory becomes less linear.

Finally, a classification model using the four predicted vocabulary ages as features was built on the 147 children with full available data and cross-validated as described in Section 3.2.

Implementation

All of the models were created and evaluated in the R programming language. For (nested) cross validation, the R package caret (Kuhn et al., 2018) was used. For both SVM and SVR, e1071 (Meyer et al., 2018) was used.

4 Results

4.1 Predicting Family Risk

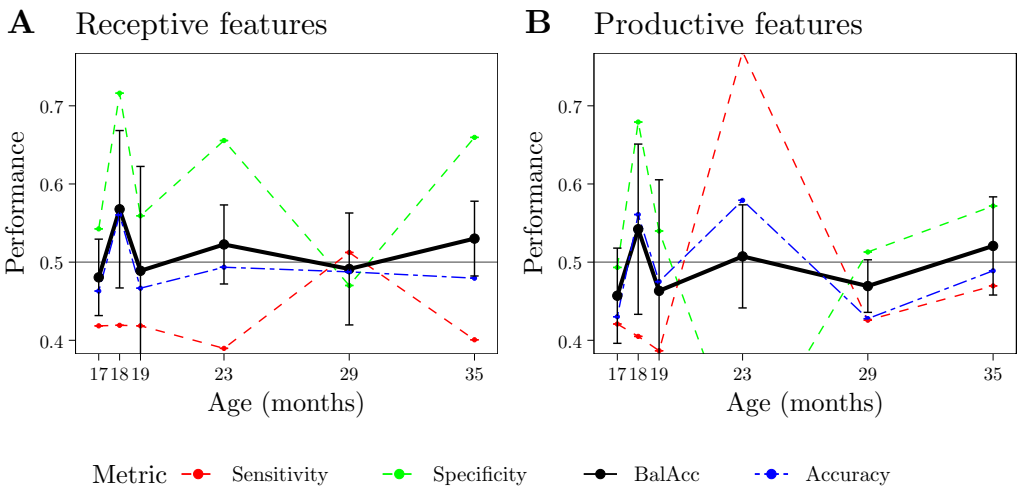


Figure 4.1: **Cross validated performance of FR-trained models**
Four different performance measures of all different crossvalidated FR-predictive models trained with class weights, on either receptive (A) or productive (B) vocabulary. Sensitivity, specificity, balanced accuracy and accuracy are shown. The standard deviation (between outer folds) for balanced accuracy is also included. The horizontal line shows 50%, the minimum performance a classifier can achieve.

Linear SVMs were trained to predict FR status. All models performed poorly, with balanced accuracies ranging from 42% to a maximum of 55% in the 18 month old group. Their AUC measures ranged from a little above 0.5 (baseline classifier) to 0.64.

When trained on the balanced additional data described in Table 3.3, results did not improve. Performance tables of the models trained on the additional data and all full results can be found in Appendix A, but a summary can be seen in Figure 4.1.

The best performing models were the REC and PROD models at 18 months, yielding 56% and 54% balanced accuracies respectively. For the models trained on the original data, balanced accuracies were always 53% or less. The 23 month old PROD model with class weights was the only one with a 70%+ sensitivity, and a 57% accuracy, but its specificity was low (45%). The best model according to the AUC measure was PROD at 19 months at 0.64. Still, its balanced accuracy was only 46%. Although many classifiers had medium AUC scores, they apparently failed to eventually pick suitable thresholds. The AUC measures per age group can be seen in Figure 4.2.

4.2 Predicting Dyslexia

SVMs were trained to predict dyslexia status. Both REC and PROD models performed poorly, with balanced accuracies ranging from 42% to 58%.

The most stable cross validated model was the oversampled PROD model at 23 months, with a

balanced accuracy of 0.58, and mean specificity/sensitivity of 49% and 67%. The AUC for this model was 0.64. This is the maximum balanced accuracy for all models, and the maximum AUC for the oversampled models. The PROD model for the same age group with weights had a 50% balanced accuracy (both sensitivity, specificity and accuracy around 50%) but a relatively high AUC of 0.66.

The AUC scores for the dyslexia models were comparable to those of the FR-predictors, ranging from 0.56 to 0.66. The AUC scores for different ages can be seen in Figure 4.3. Full results can be found in the Appendix in Table A.6.

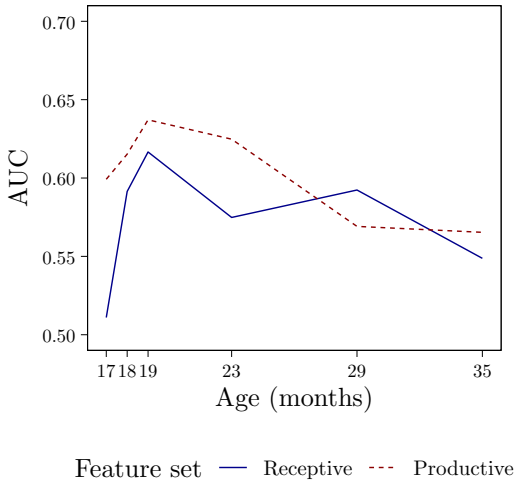


Figure 4.2: **AUC measures for FR-predictors**
AUC performance measure for FR-predicting models with class weights, for every age group.

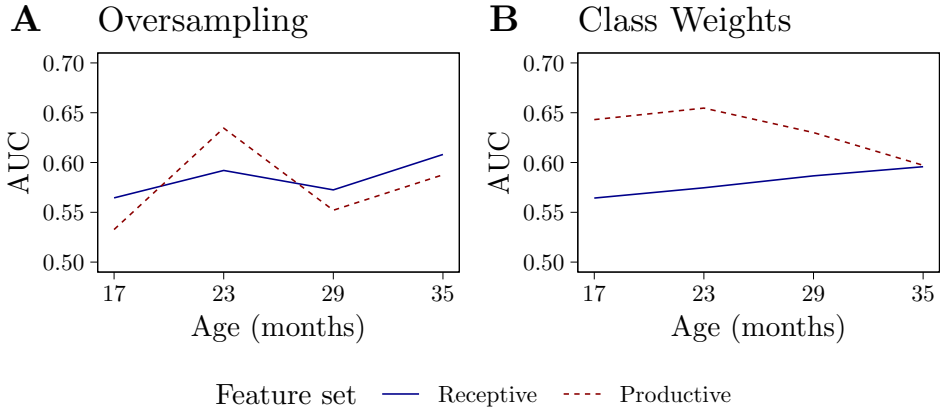


Figure 4.3: **AUC measures for all dyslexia predictors**
AUC measures for all crossvalidated models trained either with (A) oversampling or (B) class weights to correct for imbalance.

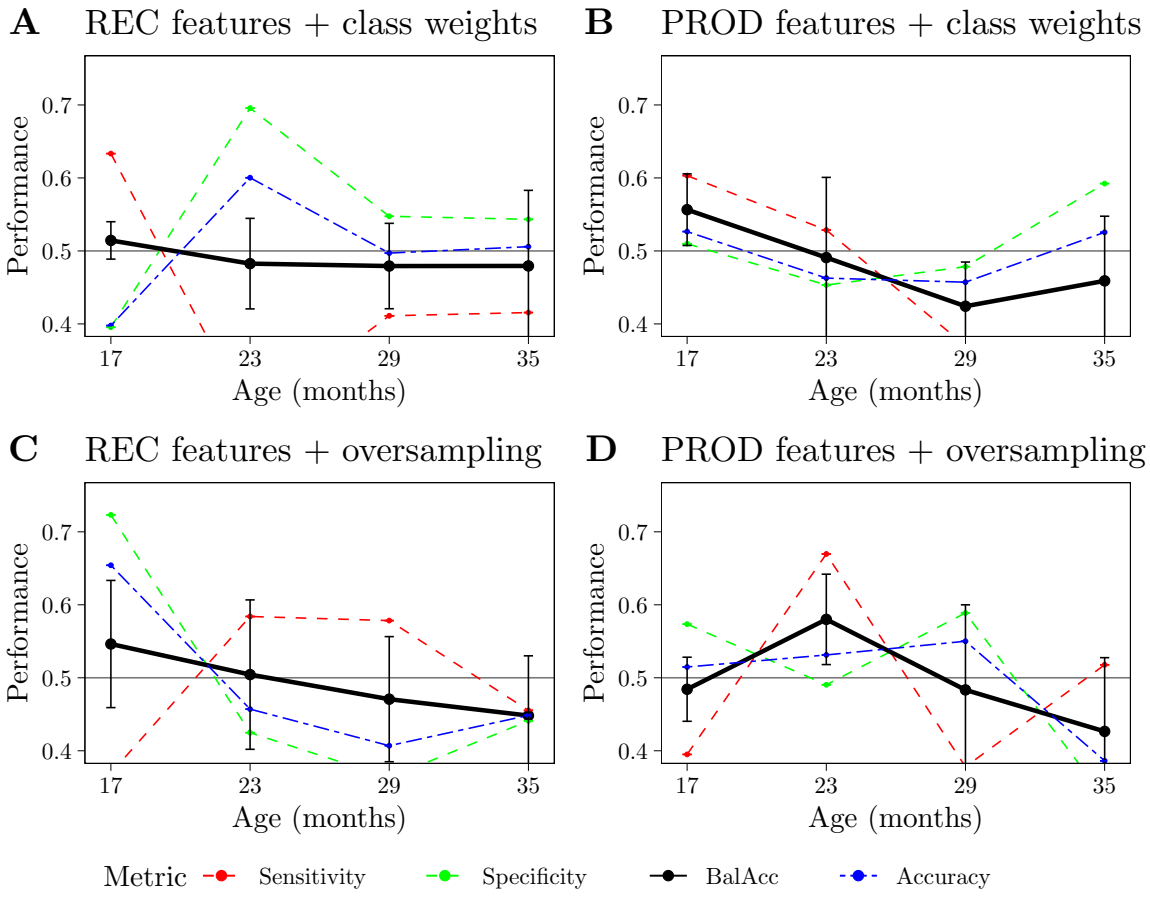


Figure 4.4: **Cross validated performance of all dyslexia-trained models**
REC or PROD signifies the feature set trained on, weights or sampling signifies which data balancing technique was used. Shown in the graphs are four different performance metrics: sensitivity, specificity, balanced accuracy and accuracy. The standard deviation (between outer folds), for balanced accuracy, our main performance metric, is also included. The horizontal line shows 50%, the minimum performance a classifier can achieve.

4.3 Predicting Age

Goodness of fit

Real Data

Receptive linear feature models provide a good fit on the train data (crossvalidated $R^2 = 0.645$). The final receptive model has an MAE of 2.387 on the control data, and generalize well when tested on the FR-ND group (MAE = 2.754) and on the FR-D group (MAE = 2.977). The PROD linear model performed slightly better, with an R^2 of 0.686. All mean absolute errors ranged from 2.3 to 3 months, which is a 17% error given the total age span[1]. The linear SVR models also generalize well when tested on the unseen data of 18 and 19 months, with similar MAE and RMSE scores. The PROD model did generalize slightly better. Results for all linear REC and PROD models can be seen in Table 4.1.

	Real Data			Generated Data		
Group	Mean error	RMSE	MAE	Mean error	RMSE	MAE
Control	-0.252	3.067	2.387	-2.904	4.410	3.506
FR-ND	-0.325	3.517	2.754	-3.099	4.994	3.892
FR-D	-1.244	3.911	2.977	-4.202	5.957	4.934
TD-Val	1.386	2.995	2.246	-3.948	5.293	4.615
FR-Val	0.926	2.631	1.914	-4.271	5.262	4.598

(a) Goodness of fit for REC Models

	Real Data			Generated Data		
Group	Mean error	RMSE	MAE	Mean error	RMSE	MAE
Control	-0.091	3.014	2.415	-0.144	3.982	3.255
FR-ND	-0.062	3.070	2.520	-0.082	3.917	3.248
FR-D	-1.237	3.614	2.819	-1.590	4.414	3.530
TD-Val	1.612	2.544	1.854	0.573	4.121	3.340
FR-Val	1.131	1.935	1.377	0.075	3.534	2.682

(b) Goodness of fit for PROD Models

Table 4.1: **Goodness of fit for REC and PROD linear models**
Age regression models were trained on only real data (left) or included generated data (right). The model was trained on the control group (first row in every table) and applied to all other groups. The last two rows, TD-Val and FR-Val are from the additional dataset for 18 to 20 months, where dyslexia status is unknown.

Adding nonlinear features to account for the nonlinear trajectory of vocabulary did not improve performance. For the full performance results of the nonlinear models, see Appendix B.

Generated Data

Adding generated data for 8 and 44 months did not improve performance either, but instead decreased it. The fit on real train data was similar (R^2=0.643 for REC-model, and R^2 =0.675 for PROD-model). As seen in Table 4.2, the MAE increased for almost all models, but especially for the linear models and the receptive models. Although the mean error decreased,tThere was more variance in the predictions, resulting in a higher MAE. Adding fewer data points did decrease variance, but those models did not outperform the models trained on real data. For full results see Appendix B1.

Effect Sizes

There was no significant difference between the Control and the FR-ND when it comes to vocabulary age. All models fit FR-ND about as well as they fit the training data. However, there were significant

[1]MAE / span = 3/(35-17)=0.167.

	Control	FR-ND	FR-DD	TD-Val	FR-Val
REC Linear	-1.119	-1.139	-1.957	-2.369	-2.684
PROD Linear	-0.840	-0.729	-0.711	-1.486	-1.306
REC Nonlinear	-0.761	-0.900	-1.620	-1.872	-2.101
PROD Nonlinear	-0.708	-0.609	-0.629	-1.505	-1.339

Table 4.2: Decline of the model performance when adding generated data. Shown in this table is $\text{MAE}_{\text{real}} - \text{MAE}_{\text{generated}}$ per feature combination and group. Nonlinear models suffer less from the added data, but still perform worse than models trained on only real data.

differences between the vocabulary ages of FR-ND and FR-D subjects, especially in the PROD trained models. Hedges' g ranged from low (0.2 for 35 month REC) to medium (0.57 for 29 months PROD) for the linear models with real data. Both the nonlinear models and the models with generated data had similar effect sizes between these two groups[2]. In Figure 4.6, the vocabulary age gap for the three groups is visualized for the best performing model. All the number differences can be seen in Table 4.3. A detailed density plot for the PROD models with linear features is available in the appendix in Figure B.2.

The biggest effect size was found in the 29 month old PROD predictions with real data, where dyslexics were estimated 1.9 months younger than their nondyslexic peers[3] ($p = 0.004$). The 23 month olds of this model were around 1.4 months behind ($p = 0.017$). The best REC effect size, again of 29 month old children, had a similar difference in months, estimating dyslexics 1.5 years younger ($p = 0.022$).

	Real Data			Generated Data		
Age	Hedges' g	Gap	Welch's t (p)	Hedges' g	Gap	Welch's t (p)
17	0.275	-0.555	1.475 ($p = 0.144$)	0.111	-0.302	0.589 ($p = 0.558$)
23	0.298	-1.027	1.531 ($p = 0.130$)	0.323	-1.252	1.643 ($p = 0.104$)
29	0.428	-1.504	2.334 ($p = 0.022$)	0.454	-1.945	2.477 ($p = 0.015$)
35	0.199	-0.557	1.042 ($p = 0.301$)	0.202	-0.814	1.096 ($p = 0.276$)

(a) Receptive

	Real Data			Generated Data		
Age	Hedges' g	Gap	Welch	Hedges' g	Gap	Welch's t (p)
17	0.319	-0.226	2.006 ($p = 0.047$)	0.468	-1.341	2.598 ($p = 0.011$)
23	0.433	-1.438	2.427 ($p = 0.017$)	0.373	-1.611	1.997 ($p = 0.049$)
29	0.565	-1.881	2.996 ($p = 0.004$)	0.498	-2.092	2.724 ($p = 0.008$)
35	0.404	-1.039	2.031 ($p = 0.046$)	0.392	-1.670	2.016 ($p = 0.047$)

(b) Productive

Table 4.3: Effect sizes between FR-ND and FR-D for all linear models, trained on either only real data (left) or generated data (right), on (a) receptive features or (b) productive features. The "Gap" column refers to the mean difference in months: FR-D − FR-ND.

Although the effect sizes were overall smaller for the youngest and oldest age groups, adding generated data to the productive models sometimes increased the differences between FR-ND and FR-D. This change can also be seen comparing Figure 4.6A and B, on page 24. For the linear PROD models, the effect size increased from an age difference of 0.2 months ($p = 0.04$) to 1.3 months ($p = 0.01$) for dyslexic children at 18 months of age. None of the REC models had this outcome.

[2]The results for models with nonlinear features can again be seen in the appendices, Table B.3.

[3]Note that this difference is relative to other children and not to actual age. As seen in Figure 4.6A, all 35 month olds are predicted to be younger than 35 months, but this is the fault of the model (regression to the mean) rather than the fault of the children's vocabulary size.

Predictive power of vocabulary age

The calculated AUC scores for different vocabulary age thresholds were all between 0.57 and 0.64 for the productive models, and between 0.45 and 0.65 for the receptive models. The AUC scores were lower for the youngest and oldest group, even after adding generated data. These AUC scores are visualized in Figure 4.5, and the numbers can be found in Table B.13 in the appendices.

Finally, SVM models were trained with the four vocabulary ages as features and crossvalidated for the original models (linear features, real data). These classification results are not better than the original, one-age classification models in section 4.2. The REC model had a balanced accuracy of 55% and an AUC of 0.62, and the PROD model a balanced accuracy of 47% and an AUC of 0.63.

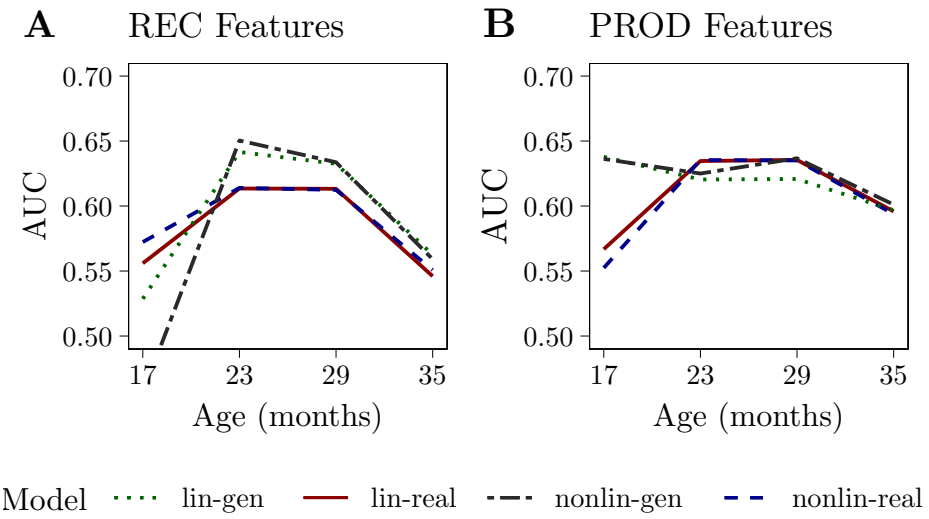


Figure 4.5: **How predictive is vocabulary age of dyslexia?**
AUC values for different age thresholds between the minimum and maximum predicted age, given a certain age group. All different models return similar AUC values for the same age groups.

Feature weights

In order to predict vocabulary age, a weight was assigned to every feature in the 22 word categories by the SVR algorithm. The feature weights for the best performing model (the linear PROD model trained on real data) are visualized in Figure 4.7 on page 25. The categories "toys", "verbs" and "helping verbs" have the largest positive weights, contributing positively to the vocabulary age outcome. Some word categories, such as "people words" and "animal names" have negative weights, although their absolute value is not as large as the biggest positive weights. Weights for other models varied. Two other weight vectors can be seen in Figure B.3.

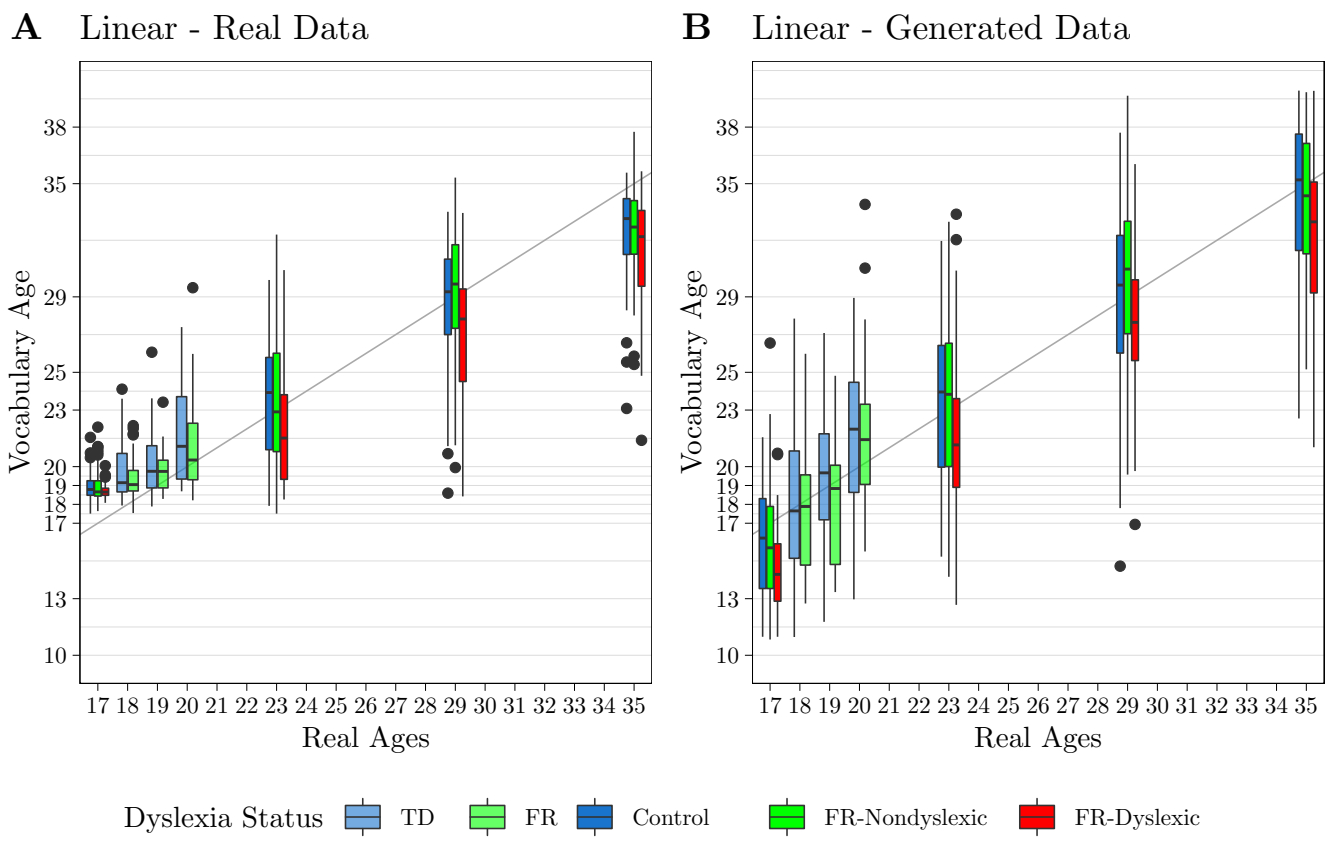


Figure 4.6: Visualized performance of two PROD-trained models, trained (A) wihout or (B) with generated data. The hinges cover the 25th to the 75th percentile of the predictions, the lines show another 1.5 interquartile range on top of that. Outliers are presented as dots. The diagonal line shows $x = y$, predictions around this line are close to the actual age. For the TD and FR groups at 18 to 20 months, dyslexia status is unknown, but the models perform similarly on this data. The model with generated data (B) has less regression to the mean, but has a wider range of predictions, especially for the younger groups. There is more variance and the fit is therefore worse. However, the dyslexic group at 17 months is more clearly behind the nondyslexic groups.

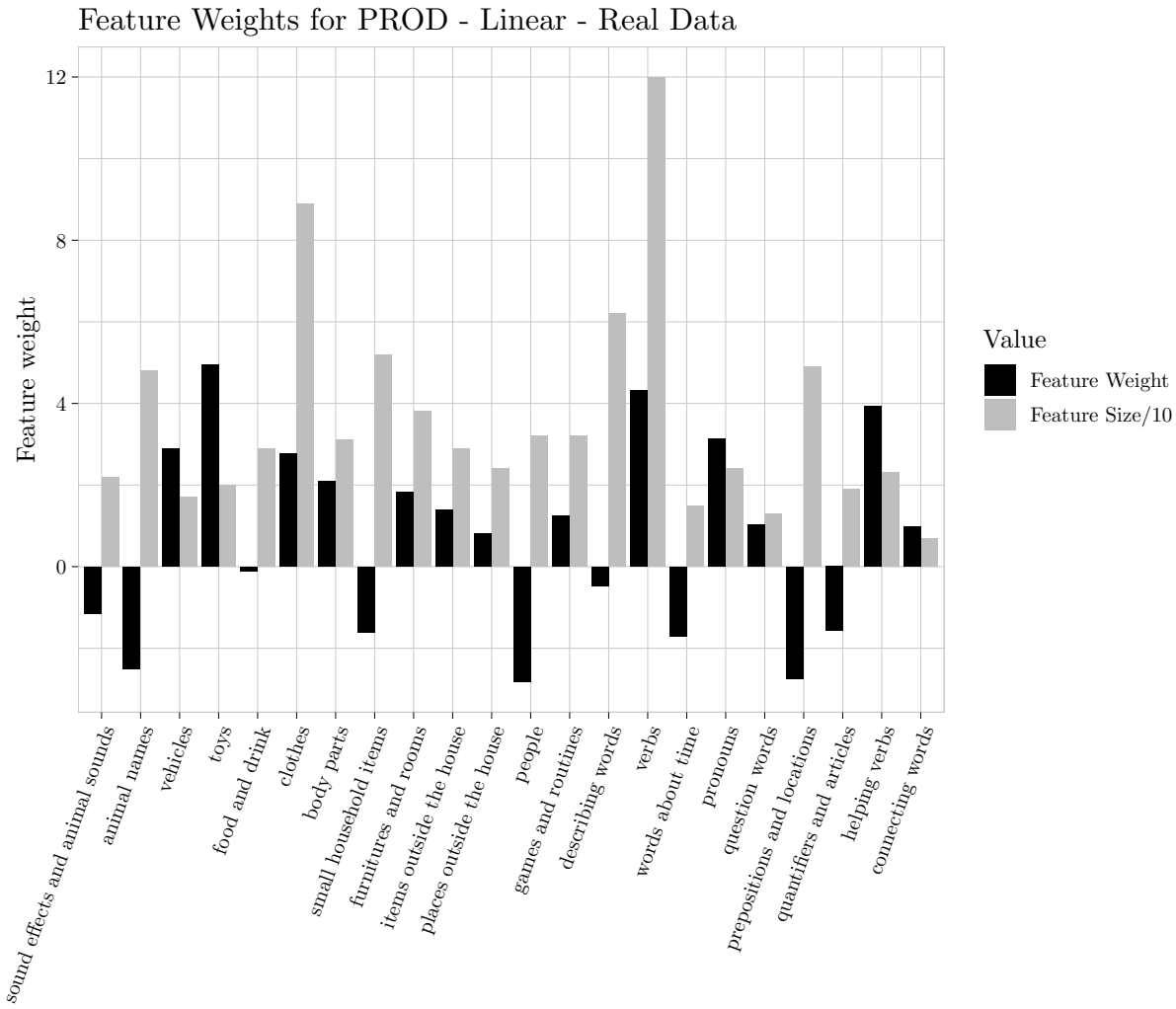


Figure 4.7: **SVR feature weights for the best performing age model (PROD - Lin - Real data)**
Shown in gray to the right of every feature weight is the size (scaled by 10) of this word category. The "verbs" category, for instance, contains 120 words. A large positive weight means the category contributes a lot to vocabulary age. On the other side, knowing many words in a negatively weighted category means the age model predicts you to be younger.

5 Discussion

5.1 Classification: Predicting Dyslexia

Linear support vector classifiers were trained to predict (familial risk of) dyslexia individually. Initially, familial risk of dyslexia was predicted in order to reproduce the original study. At first, only the complete data (subjects with known dyslexia status) was used to train the models. Then, additional data from (Chen et al., 2017) was used for more balanced results. Neither of the approaches yielded good results. Balanced accuracies ranged from 42% to 55%. Although the AUC measure was higher for the 19 to 29 month olds (the groups in the middle of the vocabulary spurt), this was not necessarily the case for the balanced accuracy or accuracy metric. When sensitivity and specificity were not both approximately 50%, they were very skewed (30%-70%). Models with a high sensitivity have a higher accuracy, because the majority group is at FR. These results are not very reliable. Even though the plotted AUC measures in Figure 4.2 seem to mimic the low-high-low performance (AUC and balanced accuracy being higher in the middle age groups) of the models reported in the original study, results are worse. Especially the sensitivity/specificity trade-off is worse. Whereas the best model in the original paper had a sensitivity of 65% and a specificity of 72%, the new cross-validated models struggled to get both metrics above 50% at the same time.

The main methodological difference with (Chen et al., 2017) is model selection. The original models were assessed with LOOCV, and C was not tuned in a nested loop. This time, the best C for the folds of every model varied wildly, which may have led to higher variance in the final assessment. The small dataset did not really allow for much model selection, for instance, $K = 10$ was not desired because of the small sample size. Using 5-fold CV over LOOCV has probably increased bias, giving all models less training data and a worse fit. In (Chen et al., 2017), predictive perormance peaked at 19 months, which may have been a criticical point in vocabulary development. Unfortunately, data for this age group was not available.

Dyslexia

As for the prediction of dyslexia, results were comparable. The dyslexia-trained models usually suffered from a low sensitivity in comparison to the specificity, meaning the detection rate of dyslexia was low. Whereas the balanced accuracy was always low (42%-58%), AUC scores were relatively better (0.56-0.66).

So far, the relatively high AUC measures have given the models some credit: there is some predictive power in the features, but the models just cannot find the best thresholds[1].

However, for highly imbalanced data, the AUC value might be too optimistic (He and Garcia, 2009). Shown in Figure 5.1 are some typical ROC curves of a model validated on one outer fold. Because there are so little positive subjects (as described in chapter 3), the ROC curve is bumpy, resulting in a large but unrealistic AUC. Increasing the number of positive subjects with $K = 5$ gave a more realistic number for how well the final model performed.

A feature that was not used was the FR-status of children. Including FR as a feature improved all measures (specificity, sensitivity, AUC) for the prediction of dyslexia in (Thompson et al., 2015). Since this dyslexic group was a subset of the FR group, using FR as a feature would be unfair. Moreover, it might be preferable to predict dyslexia just on the basis of vocabulary and not FR.

[1] Hand-picking better thresholds increased balanced accuracy to up to 68% for the 23 month old PROD model (see Table A.8 in the appendices). Still, this is the best possible performance we can force out of the model, and it is still a medium balanced accuracy.

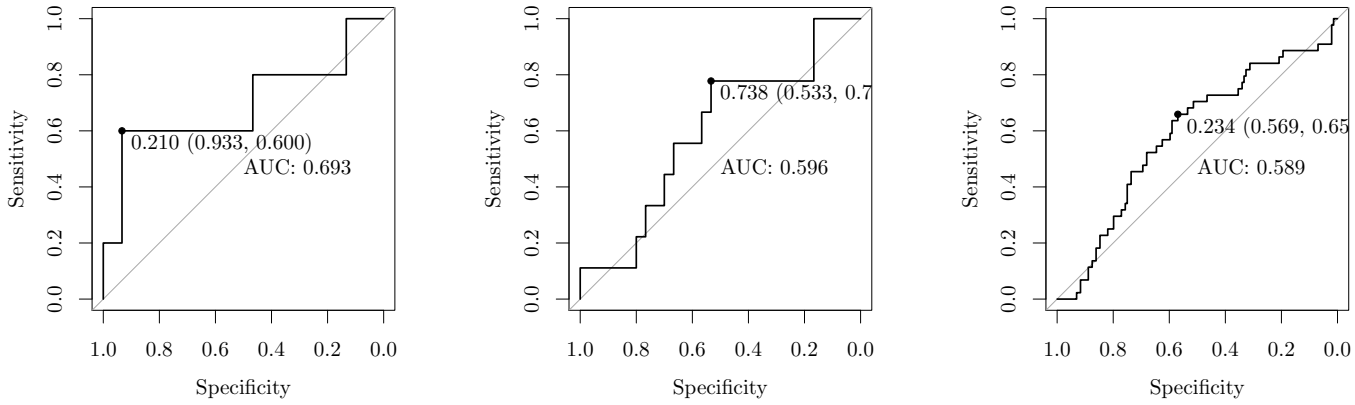


Figure 5.1: **ROC curves for imbalanced data**.
Two ROC curves for one outer fold with $K = 10$ and $K = 5$. and final model after crossvalidation. Every "step" in sensitivity corresponds to about one more dyslexic subject being classified correctly. Curves plotted from the 29 month productive model with class weights. Shown on curve is best suggested threshold.

Hyperparameters had to be tuned, which is the reason nested crossvalidation was used. An inner fold during the training phase would be of size $190/25 \approx 8$. Based on these eight test subjects, some C was chosen. Even with good resampling strategies and fair assessment metrics, there may have been too little data for reliable, crossvalidated results in the first place.

Lastly, a linear classifier may not be the right solution for this problem. All the trained classification models are insensitive to any nonlinear changes[2]. A linear SVM can in some cases be more successful than one with a nonlinear kernel[3]. Linear models are also less likely to overfit, and their feature weights are relatively easy to interpret. For these reasons, linear models were used.

5.2 Predicting Age

Linear support vector regression was used to predict vocabulary age. The models fit well on the train data of control subjects, and performed similarly on the FR-ND group, but slightly worse on the FR-D group. Dyslexics were, on the average, predicted up to two months younger than non-dyslexics. The largest effect size was found in the 29 month old group, with $g = 0.57$ and $t = 2.99$. Nonlinear features were added to give the models space to fit a more sigmoidal curve, but performance did not increase.

Generated data was added to give the models more "insight" into the vocabularies of children outside of the recorded age frame. The first models trained on real data showed regression toward the mean, i.e., the youngest age group was constantly predicted to be slightly older, and the eldest age group was constantly predicted to be slightly younger. Adding two outer age groups could help in counteracting this phenomenon. This countermeasure helped, but overall performance decreased. Models did account for the outermost real age groups (17 and 35 months), but had to give up some performance in the relevant middle groups. Variance increased for all age groups. This method of adding data may not have been nuanced enough[4]. The nonlinear models seemed to deal slightly better with the added data than the linear ones, but not consistently so. Adding fewer data points

[2]It can also be insensitive to possible heterogeneity. In the Appendix in Figure B.2, a small group of dyslexic subjects is ahead of the other dyslexic subjects. Nondyslexic subjects seem to have just one "hump".
[3]For instance, in (Kassraian-Fard et al., 2016).
[4]With different age groups on either side (5 and 47 months, or 11 and 41 months), performance did not change.

(20 instead of 60 generated datapoints) yielded better results for the PROD generated models, but not for the REC generated models. Again, these results were not better than the original PROD real models, but very similar. Differences between groups (FR-ND and FR-D) were slightly better, but AUC scores were not.

An interesting side effect of adding generated data was that group differences increased for some age groups. Whereas there was almost no difference between the predictions for FR-ND and FR-D at 17 months (FR-D predicted to be 0.2 months younger, $t = 2$), the same model trained on generated data gave a difference of 1.3 months ($t = 2.6$). This effect can also be seen in Figure B.2.

These models were trained on control subjects of all different age groups and predicts age fairly well. They have knowledge of all ages opposed to the classification models, which were trained on one age group only. Still, there was too much overlap between the two groups to set a vocabulary age-threshold separating dyslexics from non-dyslexics: AUC scores ranged from 0.55 to 0.65.

Finally, the weights of the best performing model were mostly positive. This makes intuitive sense: a larger vocabulary should not contribute to a younger vocabulary age. The biggest negative weight was the "people" category, which includes words such as "mom" and "dad". These simple words might be the reason some feature weights are negative. On the other side, a large positive feature weight might indicate a category containing more complex words not understood by younger children. The largest feature weight was that of the "toys" category. "Verbs" was the second largest feature weight. This is also the biggest category with 120 different words: it is less prone to noise than smaller categories. This reliability could explain large feature weights for larger categories in the weight vector in Figure 4.7.

It should however be noted that feature weights differed for similarly performing models: these specific weights should not be overinterpreted. For instance, both the REC model with linear features trained on real data and the PROD model with linear features trained on generated data assigned smaller weights to the "verbs" category. The "people" category had a large positive weight in the latter model. These weights can be found in the appendices in Figure B.3.

5.3 Further research

This study differed from previous research in its use of methods from the field of machine learning. Of course, only some ML techniques could be used. Vocabulary had no predictive power when using linear SVMs, but different (nonlinear) modelling techniques could result in better performance.

In addition to using nonlinear methods, ensemble methods could be employed to combine multiple weak learners (with low to medium predictive power) into a stronger learner. Ensemble methods, such a boosting, can decrease bias and variance by letting the weak learners "vote" for each prediction.

Another approach might be one-classification. This may be used for highly imbalanced data. The "classifiers" are trained on mainly, or only one class. For this problem, say, non-dyslexics. These models should learn the concept of a typical vocabulary size. Rather than differentiating between positive and negative class instances, they recognize whether a given instance is in line with the learned concept (He and Garcia, 2009). A dyslexic subject would be considered an anomaly, and not belong to this group. For some problems, this approach of outlier detection can be more successful than conventional learning approaches.

Currently, most of the research efforts in imbalanced learning focus on specific algorithms and/or case studies[5]; only a limited amount of theoretical understanding on the principles and consequences of this problem have been addressed (He and Garcia, 2009). Fields that suffer a data imbalance, such as psychiatry and psychology, might benefit from more theoretical research on this subject.

[5] Admittedly, this is one of them.

As for prediction of age, the age prediction models that were not explicitly trained to classify dyslexia could do so implicitly by predicting a lower vocabulary age for dyslxic subjects. This vocabulary age gap could be explored more in-depth. In the same line, using vocabulary data from multiple ages could yield better results, since early vocabulary trajectory differs in dyslexic children (van Viersen et al., 2017).

It should be kept in mind that the link between early vocabulary and later dyslexia is weak. Good predictors might not be achievable, even when using complex methods.

6 Conclusion

This study examined whether dyslexia could be predicted on an individual level using infant vocabulary at 17 to 35 months. Whereas earlier work was largely correlational, the aim was to use techniques from the field of machine learning, such as cross validation, to build and assess models. In line with previous results (Thompson et al., 2015; Duff et al., 2015b,a), early vocabulary was not a good predictor of dyslexia. All models, trained either on receptive of productive vocabulary, performed poorly. The small sample size and class imbalance made it difficult to tune the models. Vocabulary was a good predictor of age. Even though the age models were not trained to predict dyslexia, dyslexics were significantly behind nondyslexics when it came to vocabulary age. This gap was not large enough to individually predict dyslexia, but future research could examine this gap further. Taking vocabulary trajectory into account, as the age prediction models did, rather than training on one age group, might provide more valuable results. The imbalance of data should be kept in mind when using machine learning in the field of psychology and psychiatry, both when choosing methods and reporting model performance.

Bibliography

Cawley, G. and L. C. Talbot, N. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

Chen, A., Wijnen, F., Koster, C., and Schnack, H. (2017). Individualized early prediction of familial risk of dyslexia: A study of infant vocabulary development. *Frontiers in Psychology*, 8:156.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Duff, F. J., Nation, K., Plunkett, K., and Bishop, D. (2015a). Early prediction of language and literacy problems: is 18 months too early? *PeerJ*, 3:e1098.

Duff, F. J., Reen, G., Plunkett, K., and Nation, K. (2015b). Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry*, 56(8):848–856.

Hamilton, A., Plunkett, K., and Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of child language*, 27(3):689–705.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H., and Wenderoth, N. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in Psychiatry*, 7:177.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2018). *caret: Classification and Regression Training*. R package version 6.0-80.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-0.

Schnack, H. G. and Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, 7:50.

Schnack, H. G., van Haren, N. E., Nieuwenhuis, M., Hulshoff Pol, H. E., Cahn, W., and Kahn, R. S. (2016). Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. *American Journal of Psychiatry*, 173(6):607–616. PMID: 26917166.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., and Snowling, M. J. (2015). Developmental dyslexia: predicting individual risk. *Journal of Child Psychology and Psychiatry*, 56(9):976–987.

van Viersen, S., de Bree, E. H., Verdam, M., Krikhaar, E., Maassen, B., van der Leij, A., and de Jong, P. F. (2017). Delayed early vocabulary development in children at family risk of dyslexia. *Journal of Speech, Language, and Hearing Research*, 60(4):937–949.

Zink, I. and Dejaegere, M. (2002). *N-CDI Lijsten voor Communicatieve Ontwikkeling.* Leusden: Acco.

A Predicting Dyslexia

Reproduced Results

Predicting FR with nested crossvalidation as described in Section 3.2. Used were either class weights (Table A.1) or oversampling (Table A.2). Since the additional data was already balanced, no balancing methods were needed for the results in Table A.3. This also goes for the 18 and 19 month old group.

Age	Feature Set	Balanced Accuracy	AUC	Sensitivity	Specificity	Accuracy
17	REC	0.481	0.511	0.418	0.543	0.463
18	REC	0.568	0.591	0.419	0.716	0.560
19	REC	0.489	0.617	0.419	0.559	0.467
23	REC	0.523	0.575	0.390	0.656	0.493
29	REC	0.491	0.592	0.512	0.470	0.488
35	REC	0.530	0.549	0.401	0.660	0.479
17	PROD	0.457	0.599	0.421	0.493	0.430
18	PROD	0.542	0.615	0.405	0.679	0.561
19	PROD	0.463	0.637	0.387	0.540	0.475
23	PROD	0.507	0.625	0.769	0.245	0.579
29	PROD	0.469	0.569	0.426	0.513	0.428
35	PROD	0.521	0.565	0.470	0.572	0.489

Table A.1: Predicting FR: reproduced results, models trained with class weights

Age	Feature Set	Balanced Accuracy	AUC	Sensitivity	Specificity	Accuracy
17	REC	0.423	0.593	0.611	0.235	0.493
23	REC	0.498	0.563	0.540	0.455	0.511
29	REC	0.518	0.604	0.582	0.453	0.543
35	REC	0.453	0.577	0.606	0.300	0.515
17	PROD	0.531	0.582	0.610	0.452	0.559
23	PROD	0.480	0.528	0.524	0.437	0.500
29	PROD	0.465	0.605	0.451	0.480	0.457
35	PROD	0.502	0.595	0.486	0.519	0.484

Table A.2: Predicting FR: reproduced results, models trained with oversampling

Age	Feature Set	Balanced Accuracy	AUC	Sensitivity	Specificity	Accuracy
17	REC	0.495	0.543	0.390	0.600	0.482
23	REC	0.476	0.554	0.688	0.263	0.488
29	REC	0.491	0.520	0.190	0.792	0.497
35	REC	0.610	0.570	0.703	0.517	0.613
17	PROD	0.489	0.548	0.243	0.735	0.503
23	PROD	0.448	0.615	0.439	0.457	0.468
29	PROD	0.462	0.632	0.383	0.542	0.461
35	PROD	0.484	0.561	0.643	0.326	0.500

Table A.3: Predicting FR: reproduced results for balanced additional data

Dyslexia Prediction Results

First, the built FR-models were used to predict dyslexia. The results can be seen in Table A.4.

Age	Feature Set	AUC
17	REC	0.601
23	REC	0.536
29	REC	0.581
35	REC	0.557
17	PROD	0.576
23	PROD	0.605
29	PROD	0.505
35	PROD	0.581

Table A.4: **Predicting Dyslexia with FR-trained models**
Cross-validated AUC score for testing final outer fold models of FR-trained models on DYS. BalAcc is not included since it is always 0.5 (sens=0, spec=1). AUC scores are slightly lower than true DYS-trained models, but not extremely low. The features that are predictive of FR are also somewhat predictive of dyslexia.

Then, dyslexia models were trained as describe in Section 3.2.

Age	Feature Set	Balanced Accuracy	AUC	Sensitivity	Specificity	Accuracy
17	REC	0.514	0.564	0.633	0.395	0.398
23	REC	0.483	0.575	0.270	0.696	0.600
29	REC	0.479	0.587	0.411	0.547	0.497
35	REC	0.479	0.596	0.416	0.543	0.506
17	PROD	0.557	0.643	0.603	0.510	0.527
23	PROD	0.491	0.655	0.529	0.453	0.463
29	PROD	0.424	0.630	0.370	0.478	0.457
35	PROD	0.459	0.597	0.326	0.592	0.526

Table A.5: Predicting dyslexia: results for models trained with oversampling

Age	Feature Set	Balanced Accuracy	AUC	Sensitivity	Specificity	Accuracy
17	REC	0.514	0.564	0.633	0.395	0.398
23	REC	0.483	0.575	0.270	0.696	0.600
29	REC	0.479	0.587	0.411	0.547	0.497
35	REC	0.479	0.596	0.416	0.543	0.506
17	PROD	0.557	0.643	0.603	0.510	0.527
23	PROD	0.491	0.655	0.529	0.453	0.463
29	PROD	0.424	0.630	0.370	0.478	0.457
35	PROD	0.459	0.597	0.326	0.592	0.526

Table A.6: Predicting dyslexia: results for models trained with class weights

Handpicked Thresholds

Instead of using the thresholds calculated by the SVM trained on four outer folds, the best threshold was picked as seen in the ROC curve. Model performance increased slightly, but a (balanced) accuracy of 64% is still low. Picking tresholds during validation can also result in overfitting.

Age	Feature Set	Balanced Accuracy	Sensitivity	Specificity	Accuracy	AUC
17	REC	0.584	0.597	0.571	0.587	0.511
23	REC	0.629	0.627	0.632	0.618	0.575
29	REC	0.626	0.556	0.696	0.610	0.592
35	REC	0.598	0.570	0.625	0.589	0.549
17	PROD	0.655	0.667	0.644	0.664	0.599
23	PROD	0.652	0.731	0.573	0.687	0.625
29	PROD	0.590	0.560	0.620	0.574	0.569
35	PROD	0.624	0.587	0.661	0.615	0.565

Table A.7: Handpicked thresholds for outer folds (class weights) - FR

Age	Feature Set	Balanced Accuracy	Sensitivity	Specificity	Accuracy	AUC
17	REC	0.635	0.620	0.649	0.632	0.564
23	REC	0.615	0.592	0.639	0.627	0.575
29	REC	0.635	0.564	0.706	0.669	0.587
35	REC	0.651	0.629	0.673	0.667	0.596
17	PROD	0.671	0.627	0.715	0.693	0.643
23	PROD	0.677	0.694	0.661	0.664	0.655
29	PROD	0.647	0.650	0.644	0.639	0.630
35	PROD	0.667	0.632	0.702	0.687	0.597

Table A.8: Hand-picked thresholds for outer folds (class weights) - Dyslexia

B Predicting Vocabulary Age

Group results

This includes results for models trained on 44 features, including the nonlinear ones. There is little to no difference for the group results and the full results between these linear and nonlinear models.

	Real Data			Generated Data		
group	mean error	RMSE	MAE	mean error	RMSE	MAE
Control	-0.252	3.067	2.387	-2.904	4.410	3.506
FR-ND	-0.325	3.517	2.754	-3.099	4.994	3.892
FR-DD	-1.244	3.911	2.977	-4.202	5.957	4.934
TD-Val	1.386	2.995	2.246	-3.948	5.293	4.615
FR-Val	0.926	2.631	1.914	-4.271	5.262	4.598

(a) REC - Linear features

	Real Data			Generated Data		
group	mean error	RMSE	MAE	mean error	RMSE	MAE
Control	-0.194	3.084	2.410	-2.468	4.068	3.171
FR-ND	-0.238	3.506	2.761	-2.712	4.676	3.661
FR-DD	-1.193	3.860	2.949	-3.783	5.579	4.569
TD-Val	1.026	2.835	2.116	-2.866	4.678	3.989
FR-Val	0.617	2.534	1.906	-3.378	4.656	4.008

(b) REC - Nonlinear features

Table B.1: Performance for all REC trained regression models

	Real Data			Generated Data		
group	mean error	RMSE	MAE	mean error	RMSE	MAE
Control	-0.091	3.014	2.415	-0.145	3.982	3.255
FR-ND	-0.062	3.070	2.520	0.082	3.917	3.248
FR-DD	-1.237	3.614	2.819	-1.590	4.420	3.530
TD-Val	1.612	2.544	1.854	0.573	4.143	3.340
FR-Val	1.131	1.935	1.377	0.075	3.520	2.683

(a) PROD - linear features

	Real Data			Generated Data		
group	mean error	RMSE	MAE	mean error	RMSE	MAE
Control	-0.121	2.992	2.383	-0.262	3.773	3.091
FR-ND	-0.107	3.090	2.522	-0.100	3.775	3.131
FR-DD	-1.273	3.636	2.835	-1.763	4.291	3.464
TD-Val	1.548	2.507	1.828	0.486	4.044	3.333
FR-Val	1.081	1.903	1.361	-0.031	3.462	2.700

(b) PROD - Nonlinear features

Table B.2: Performace for all PROD trained regression models.

	Real Data			Generated Data		
Age	Hedges' g	Gap	Welch	Hedges' g	Gap	Welch
17	0.315	-0.673	1.679 ($p = 0.097$)	0.154	-0.415	0.811 ($p = 0.420$)
23	0.311	-1.072	1.600 ($p = 0.114$)	0.324	-1.187	1.655 ($p = 0.102$)
29	0.432	-1.497	2.353 ($p = 0.021$)	0.455	-1.839	2.480 ($p = 0.015$)
35	0.206	-0.561	1.083 ($p = 0.282$)	0.196	-0.762	1.051 ($p = 0.296$)

(a) REC - Nonlinear features

	Real Data			Generated Data		
Age	Hedges' g	Gap	Welch	Hedges' g	Gap	Welch
17	0.289	-0.207	1.826 ($p = 0.071$)	0.459	-1.294	2.547 ($p = 0.012$)
23	0.428	-1.429	2.402 ($p = 0.018$)	0.400	-1.641	2.156 ($p = 0.034$)
29	0.559	-1.883	2.977 ($p = 0.004$)	0.547	-2.108	2.991 ($p = 0.004$)
35	0.387	-1.027	1.954 ($p = 0.055$)	0.387	-1.630	2.008 ($p = 0.048$)

(b) PROD - Nonlinear features

Table B.3: Effect sizes for all nonlinear models

Full results per age group

Results for generated data were included: Real-data models could not reach the added ages. No difference in generated data was added between groups. Errors for generated ages were not included in the group results section above.

Group	Age	Mean error	RMSE	MAE
NA	8	8.249	0.124	8.249
Control	17	1.346	2.252	1.603
FR-ND	17	1.783	2.737	1.992
FR-DD	17	1.228	2.189	1.570
TD-Val	18	1.562	3.236	2.414
FR-Val	18	1.456	2.907	2.053
TD-Val	19	1.266	2.760	2.143
FR-Val	19	0.577	2.436	1.822
TD-Val	20	1.173	2.736	2.032
FR-Val	20	0.278	2.244	1.743
Control	23	0.921	2.881	2.297
FR-ND	23	0.871	3.334	2.654
FR-DD	23	-0.156	3.717	2.887
Control	29	-0.417	3.211	2.650
FR-ND	29	-0.148	3.381	2.808
FR-DD	29	-1.652	3.971	3.243
Control	35	-3.029	3.770	3.029
FR-ND	35	-3.451	4.324	3.459
FR-DD	35	-4.008	5.037	4.008
NA	44	-8.679	8.679	8.679

(a) REC - Linear - Real Data

Mean error	RMSE	MAE
1.016	1.045	1.016
-1.668	2.826	2.327
-1.655	3.205	2.651
-1.957	3.191	2.626
-3.191	4.923	4.221
-3.626	4.844	4.162
-4.629	5.602	4.980
-4.275	5.041	4.429
-4.704	5.640	4.997
-5.455	6.132	5.558
-2.193	3.477	2.976
-2.286	4.232	3.471
-3.538	5.537	4.879
-2.980	4.932	3.898
-2.836	5.010	3.866
-4.781	6.495	5.707
-4.899	5.846	4.899
-5.372	6.645	5.394
-6.186	7.423	6.186
-6.135	6.140	6.135

(b) REC - Linear - Gen. Data

Table B.4: Full results for REC - Linear features

Group	Age	Mean error	RMSE	MAE
NA	8	7.713	7.714	7.713
Control	17	1.333	2.347	1.651
FR-ND	17	1.771	2.807	2.072
FR-DD	17	1.098	2.235	1.594
TD-Val	18	1.237	3.047	2.268
FR-Val	18	1.150	2.731	1.956
TD-Val	19	0.907	2.666	2.063
FR-Val	19	0.315	2.404	1.917
TD-Val	20	0.749	2.572	1.887
FR-Val	20	-0.080	2.259	1.806
Control	23	1.056	2.964	2.385
FR-ND	23	1.066	3.389	2.699
FR-DD	23	-0.006	3.708	2.887
Control	29	-0.322	3.179	2.619
FR-ND	29	-0.034	3.334	2.776
FR-DD	29	-1.531	3.881	3.161
Control	35	-3.016	3.729	3.016
FR-ND	35	-3.405	4.245	3.405
FR-DD	35	-3.966	4.944	3.966
NA	44	-8.796	8.797	8.796

(a) REC - Nonlinear - Real data

Mean error	RMSE	MAE
0.715	0.775	0.715
-1.158	2.608	2.065
-1.139	2.935	2.425
-1.554	2.967	2.404
-2.363	4.601	3.897
-2.841	4.444	3.769
-3.417	4.855	4.222
-3.461	4.520	3.853
-3.276	4.649	3.939
-4.288	5.136	4.592
-1.592	2.971	2.519
-1.762	3.822	3.163
-2.949	4.965	4.326
-2.655	4.513	3.581
-2.524	4.640	3.700
-4.362	6.023	5.304
-4.595	5.566	4.595
-5.155	6.366	5.168
-5.917	7.166	5.917
-5.739	5.745	5.739

(b) REC - Nonlinear - Gen. Data

Table B.5: Full results for REC - Nonlinear features

Group	Age	Mean error	RMSE	MAE
NA	8	10.725	10.725	10.725
Control	17	1.922	2.063	1.922
FR-ND	17	1.940	2.104	1.940
FR-DD	17	1.714	1.757	1.714
TD-Val	18	1.796	2.346	1.798
FR-Val	18	1.293	1.642	1.326
TD-Val	19	1.173	2.279	1.535
FR-Val	19	0.769	1.323	0.943
TD-Val	20	1.686	3.074	2.255
FR-Val	20	1.173	2.746	1.875
Control	23	0.342	2.990	2.648
FR-ND	23	0.380	3.460	2.857
FR-DD	23	-1.058	3.140	2.622
Control	29	-0.212	3.191	2.473
FR-ND	29	0.250	3.075	2.535
FR-DD	29	-1.631	3.984	3.109
Control	35	-2.569	3.632	2.627
FR-ND	35	-2.525	3.375	2.696
FR-DD	35	-3.564	4.675	3.665
NA	44	-8.597	8.597	8.597

(a) PROD - Linear - Real data

Mean error	RMSE	MAE
2.992	2.994	2.992
-0.965	3.007	2.490
-1.040	3.206	2.652
-2.381	3.357	2.940
-0.091	4.034	3.211
-0.397	3.120	2.539
0.786	3.770	3.071
-0.974	3.185	2.539
1.597	4.644	3.830
1.924	4.400	3.083
0.285	3.952	3.368
0.328	4.263	3.586
-1.282	4.462	3.580
0.351	4.580	3.694
1.119	4.206	3.443
-0.973	4.402	3.496
-0.274	4.211	3.465
-0.171	3.854	3.267
-1.841	5.150	4.030
-3.015	3.017	3.015

(b) PROD - Linear - Gen. Data

Table B.6: Full results for PROD - Linear features

Group	Age	Mean error	RMSE	MAE
NA	8	10.760	10.760	10.760
Control	17	1.849	1.998	1.849
FR-ND	17	1.878	2.051	1.878
FR-DD	17	1.671	1.715	1.671
TD-Val	18	1.718	2.283	1.732
FR-Val	18	1.243	1.627	1.294
TD-Val	19	1.125	2.264	1.563
FR-Val	19	0.735	1.320	0.963
TD-Val	20	1.632	3.058	2.255
FR-Val	20	1.108	2.678	1.857
Control	23	0.265	2.970	2.618
FR-ND	23	0.291	3.467	2.863
FR-DD	23	-1.138	3.178	2.658
Control	29	-0.231	3.192	2.467
FR-ND	29	0.203	3.127	2.571
FR-DD	29	-1.680	4.014	3.154
Control	35	-2.515	3.609	2.609
FR-ND	35	-2.512	3.420	2.721
FR-DD	35	-3.539	4.704	3.684
NA	44	-8.357	8.357	8.357

(a) PROD - Nonlinear - Real data

Mean error	RMSE	MAE
2.692	2.695	2.692
-0.940	3.065	2.527
-1.121	3.190	2.652
-2.415	3.355	2.957
-0.230	4.027	3.277
-0.545	3.137	2.538
0.698	3.755	3.166
-0.943	3.124	2.612
1.604	4.327	3.593
1.767	4.245	3.079
0.379	3.746	3.168
0.435	4.088	3.423
-1.207	4.183	3.282
0.147	4.082	3.320
0.816	3.819	3.150
-1.292	4.140	3.304
-0.673	4.117	3.358
-0.589	3.897	3.261
-2.219	5.179	4.247
-2.717	2.720	2.717

(b) PROD - Nonlinear - Gen. Data

Table B.7: Full results for PROD - Nonlinear features

B.1 Adding fewer data points

Adding fewer data points (20 control subjects instead of 60) resulted in a better fit for the PROD models. In the boxplots in Figure B.1 it can be seen that variance is lower for the younger age groups. This time, the nonlinear models did not perform better. Performance for these models are comparable to the original PROD linear models, but they suffer less from regression to the mean.
The REC models performed just as bad as with the large amount of datapoints added. AUC scores for all of the models did not increase either. Full performance results and effect size results can be found in the tables on the next page.

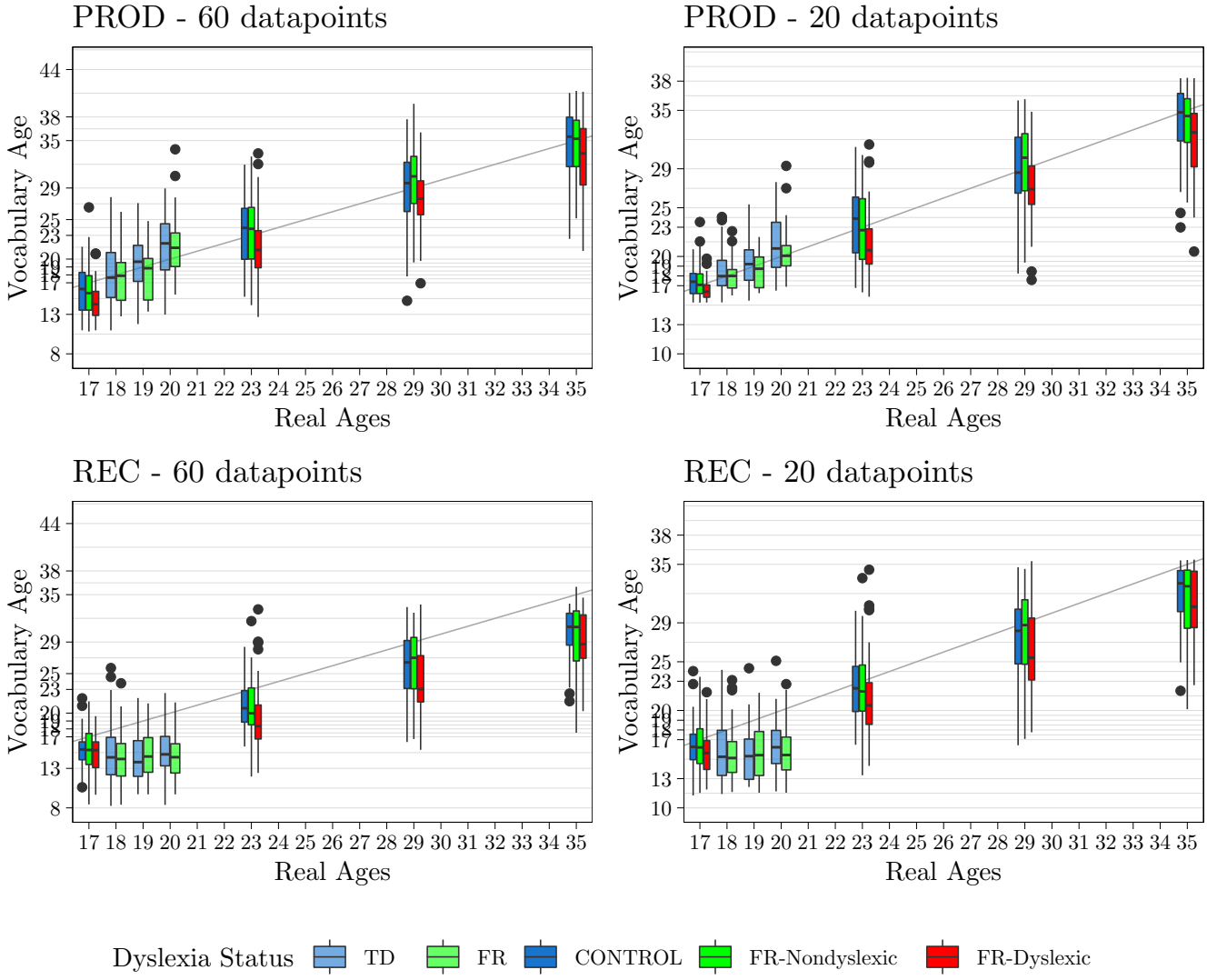


Figure B.1: **Age prediction model performance for varying amounts of data**
Visualization of age prediction models trained with linear features (REC or PROD), for varying amounts of added data. The two leftmost plots are of models trained on 60 datapoints, the two rightmost plots are of models trained on only 20 datapoints. The data was added on both sides of the original sample at 8 and 44 months.
The hinges cover the 25th to the 75th percentile of the predictions, the lines show another 1.5 interquartile range on top of that. Outliers are presented as dots. The diagonal line shows $x = y$, predictions around this line are close to the actual age.

	Linear			Nonlinear		
Group	Mean error	RMSE	MAE	Mean error	RMSE	MAE
Control	-1.395	3.586	2.704	-1.558	3.553	2.659
FR-ND	-1.523	4.090	3.127	-1.687	4.063	3.111
FR-DD	-2.654	4.847	3.887	-2.739	4.801	3.804
TD-Val	-2.796	4.109	3.571	-2.012	3.778	3.215
FR-Val	-3.148	4.206	3.707	-2.399	3.805	3.234

Table B.8: Goodness of fit for all REC-trained models with 20 extra datapoints. Models either trained with linear features (left) or nonlinear features included (right).

	Linear			Nonlinear		
Group	Mean error	RMSE	MAE	Mean error	RMSE	MAE
Control	-0.152	3.352	2.617	-0.257	3.394	2.679
FR-ND	-0.095	3.192	2.522	-0.180	3.258	2.608
FR-DD	-1.590	3.778	2.844	-1.725	3.864	2.984
TD-Val	0.753	2.717	2.083	0.543	2.891	2.301
FR-Val	0.148	2.078	1.533	-0.097	2.292	1.761

Table B.9: Goodness of fit for all PROD-trained models with 20 extra datapoints. Models either trained with linear features (left) or nonlinear features included (right).

	Linear			Nonlinear		
Age	Hedges' g	Abs	Welch	Hedges' g	Abs	Welch
17	0.262	-0.693	1.389 ($p = 0.169$)	0.204	-0.523	1.078 ($p = 0.284$)
23	0.288	-1.124	1.473 ($p = 0.145$)	0.322	-1.143	1.651 ($p = 0.103$)
29	0.438	-1.882	2.401 ($p = 0.018$)	0.441	-1.729	2.408 ($p = 0.018$)
35	0.208	-0.791	1.114 ($p = 0.268$)	0.204	-0.750	1.095 ($p = 0.277$)

Table B.10: Difference in predictions between FR-D and FR-ND children for all REC-trained models, trained including 20 generated datapoints. Models either trained with linear features (left) or nonlinear features included (right).

	Linear			Nonlinear		
Age	Hedges' g	Gap	Welch	Hedges' g	Gap	Welch
17	0.475	-0.685	2.746 ($p = 0.007$)	0.475	-0.785	2.732 ($p = 0.007$)
23	0.391	-1.417	2.135 ($p = 0.035$)	0.396	-1.478	2.162 ($p = 0.033$)
29	0.600	-2.274	3.279 ($p = 0.001$)	0.610	-2.305	3.325 ($p = 0.001$)
35	0.448	-1.516	2.294 ($p = 0.025$)	0.445	-1.546	2.279 ($p = 0.026$)

Table B.11: Difference in predictions between FR-D and FR-ND children for all PROD-trained models, trained including 20 generated datapoints. Models either trained with linear features (left) or nonlinear features included (right).

Age	17	23	29	35
REC - Linear - Gen. Data (20)	0.588	0.626	0.625	0.558
REC - Nonlinear - Gen. Data (20)	0.555	0.636	0.628	0.560
PROD - Linear - Gen. Data (20)	0.641	0.629	0.644	0.612
PROD - Nonlinear - Gen. Data (20)	0.640	0.630	0.648	0.609

Table B.12: AUC scores for all age prediction models, trained including 20 generated datapoints.

Additional results

Age	17	23	29	35
REC - Linear - Real Data	0.556	0.613	0.613	0.546
REC - Linear - Gen. Data	0.529	0.642	0.633	0.563
REC - Nonlinear - Real Data	0.572	0.614	0.612	0.551
REC - Nonlinear - Gen. Data	0.456	0.650	0.634	0.559
PROD - Linear - Real Data	0.567	0.635	0.635	0.596
PROD - Linear - Gen. Data	0.638	0.620	0.621	0.597
PROD - Nonlinear - Real Data	0.552	0.635	0.635	0.594
PROD - Nonlinear - Gen. Data	0.636	0.625	0.637	0.601

Table B.13: AUC scores with vocabulary age thresholds for all age prediction models

First, a single SVM was trained for both feature sets (PROD or REC) using the four features generated by the linear age prediction model trained on real data: vocabulary age at 17, vocabulary age at 23, vocabulary age at 29 and vocabulary age at 35. Results are reported in Table B.14. Then, the SVMs were build as before and crossvalidated, but only using the four vocabulary age features. Results were not better than the single age-group models, as can be seen in Table B.15. Both models were trained using class weights.

	BalAcc	Sensitivity	Specificity	AUC
REC	0.602	0.546	0.658	0.670
PROD	0.614	0.667	0.561	0.682

Table B.14: Training accuracy for two SVMs trained on four features: vocabulary age at 17, vocabulary age at 23, vocabulary age at 29 and vocabulary age at 35.

	BalAcc	Sensitivity	Specificity	AUC
REC	0.549	0.733	0.365	0.618
PROD	0.471	0.333	0.579	0.625

Table B.15: Performance of crossvalidated SVMs trained on four features: vocabulary age at 17, vocabulary age at 23, vocabulary age at 29 and vocabulary age at 35.

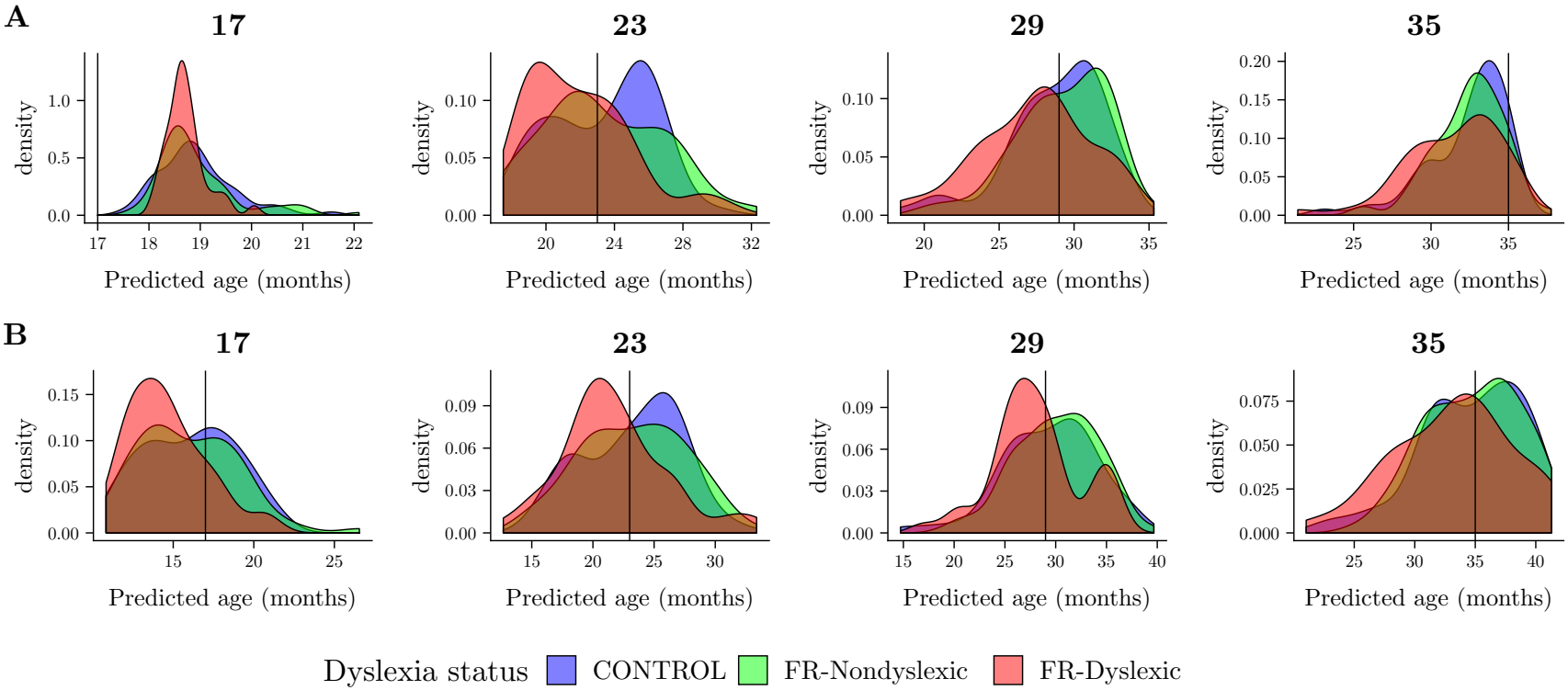


Figure B.2: Density plot of predicted ages for all thee groups. Models were trained with 22 PROD features, on (A) only real data or (B) including 60 generated datapoints. For every age group, a vertical line with the actual age is included for clarity. Adding generated data increases variance, especially at 17 months. It also increases the gap between dyslexics and nondyslexics: compare (A) 17 months and (B) 17 months. At 35 months, most dyslexic subjects seem to have "caught up". In (B) at 29 months, there is a clear group of dyslexic subjects that is catching up already.

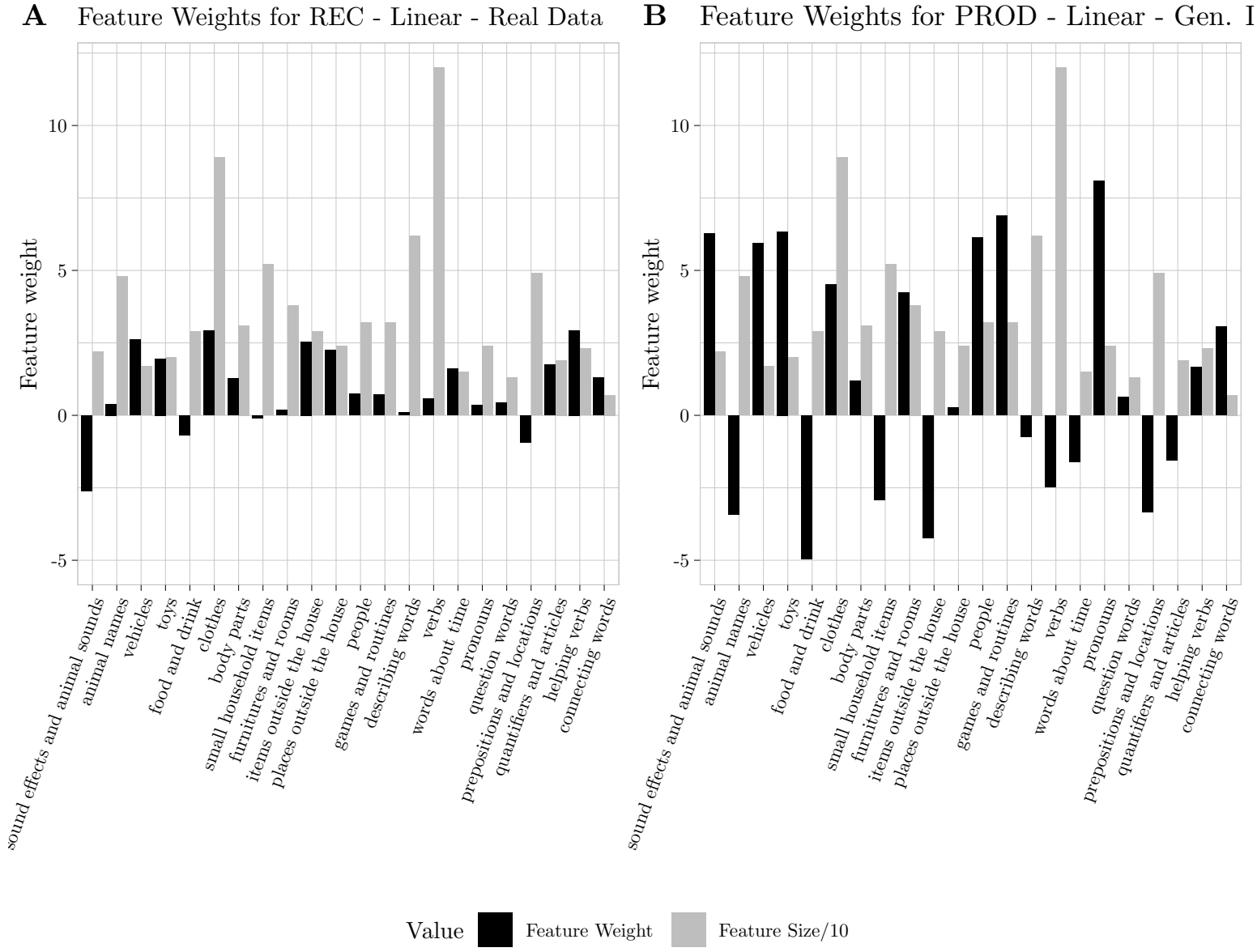


Figure B.3: **SVR feature weights for two different models**
Although their performance on unseen data is similar, these weight vectors differ from the weight vector of PROD - Linear - Real data (as shown in Figure 4.7). For instance, the REC model in (A) has almost no negative weights, and the PROD model with generated data in (B) assigns a negative weight to the "verbs" category.