A Data Science Solution to Dental Plaque

Yunbin (Matteo) Zhang
Pat Sukhum
Young Sang Choi
Seong Woo Han

Business Understanding

Dental caries, colloquially known as cavities, is the most prevalent chronic illness in America for both children and adults.[1] Cavities occur when an area on the surface of a tooth dissolves chemically due to the microbiological activity in dental plaque. Thus, by removing plaque on a regular basis through proper oral self-care, cavities are a preventable health issue.[2] What makes prevention difficult, however, is that dental plaque is initially colorless and becomes naturally visible at the onset of oral disease. Consequentially, for the 181 million Americans who do not have access to dental care, many are often unaware of their own plaque level and neglect to maintain proper oral health (See Appendix A).

For our project, we used Machine Learning to quantify the level of plaque in one's mouth to allow for an accessible and automatic dental diagnostic. Traditionally, to visualize plaque, a dentist would ask a patient to chew on a disclosing tablet which stains the plaque on the patient's teeth. The dentist would then inform the patient about the level of plaque present and educate them on proper self-care measures. With our Machine Learning algorithm, users can forgo the dentist to garner the same result. Thus, our project offers a tool for self-assessment—the user can track the amount of plaque in their mouth on a scheduled basis, encouraging the continuous oral care that prevent plaque-related issues.

On top of the social benefits from improving oral health for users, our project also has many monetization opportunities, ranging from grants and tax breaks via public health projects to funding from research labs in academia and industry. The current teledentistry market is populated by programs where dental services are administered by dental hygenists and assistants that work from temporary clinical spaces. Despite being a relatively new field, the teledentistry market is projected to grow, as demonstrated by Medi-Cal, California's state insurance program for economically disadvantaged citizens. The dental program is the first state-sponsored teledentistry project, being signed into law in 2015, and shows both a market for remote dental assessment as well as the present state of the industry. However, currently, the

---

[1] "Dental Caries (Tooth Decay)." National Institutes of Health.

[2] D.J. Fischer et al. *Dental Caries and Periodontal Conditions, in Risk Assessment and Oral Diagnostics in Clinical Dentistry*., Chapter 6.

bottleneck in teledentistry programs is in their design, as services have to be administered by a costly and limited skilled workforce in dentists and dental hygienists. The Deployment section of our writeup details a consumer-facing smartphone app that uses both our algorithm and relationships between us and dental practices to include an advertising and service commerce platform. Our app enters an untapped space in the market as it removes costly dentists from dental diagnostic, and is monetized by a referral and appointment service instead of charging for the assessment itself.[3]

Data Understanding

Our dataset consists of 468 pictures of teeth stained by disclosing tablets and labeled by NYU dental students. The images were labeled using a common technique in dental literature, where every image is scored with an integer between 0 to 5. A label of 0 corresponds to no visible plaque, and a label of 5 corresponds to the highest amount of plaque. Every image in the dataset is 200 pixels wide and 100 pixels high for a total of 20,000 pixels.

Selection bias is heavily present in the dataset due to both the small number of images and how the images were collected. Out of the 468 images in the dataset, only 8 images are labeled with a score of 3. The images were taken during a dental hackathon populated by dental and engineering students. When prompted about the lack of class 3 images in the data, the dental students who labeled the data explained the discrepancy in oral self-care habits between the two demographics. Since dental students are educated in and generally more motivated to practice proper oral hygiene than engineering students, the data is clustered in the ranges between 1-2 and 4-5.

To deal with having so few class 3 images, we used stratified sampling to split between the training and the test set to ensure that some class 3 images make it to the test set. We are not too concerned with this deficiency however, as the scores serve as arbitrary boundaries between a spectrum of stained plaque and we are measuring the amount of visible stained plaque directly.

---

[3] Sarah Sipek, "Smile Wide for the Camera." Workforce 95, no. 7: 15. Business Source Complete, EBSCOhost

Data Preparation

Since the data we have are images, we have to extract pixel information to make up our features. We know that a dental student assigns the score to an image by looking for the pink stains on the teeth so we leverage this domain knowledge as we engineer the features. Our first thought was to use raw RGB or HLS pixel values of the images as features. As there are 3 values—R, G, and B—in each pixel and there are 20,000 pixels in each image, we ended up with 60,000 features. We also tried using the average RGB and HLS values of all the pixels of an image as features (three from RGB and three from HLS for a total of six features).

Additionally, we attempted to remove noise in the data by using Computer Vision techniques such as grayscaling and blurring to see if it would improve the model. We added the raw pixel values to the model as features (20,000 features for grayscaling and 60,000 features for blurring).

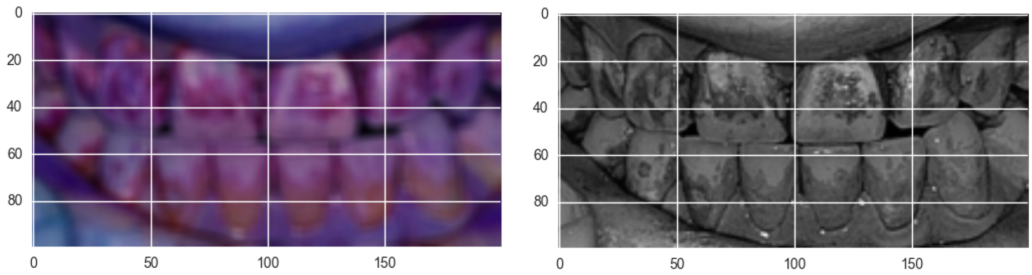


Figure 1: Blurring and grayscaling, respecitvely.

Next, we used a Computer Vision technique of masking to filter out the colors in an image that do not fit into a certain RGB range (in our case, the values that do not correspond to a certain shade of pink), leaving those pixels black. Then, we used binning to assign all the non-black pixels (the varying pink values) a value of 1 and all the other pixels to 0. Then we ran code to count the number of non-black pixels to be used as a feature.

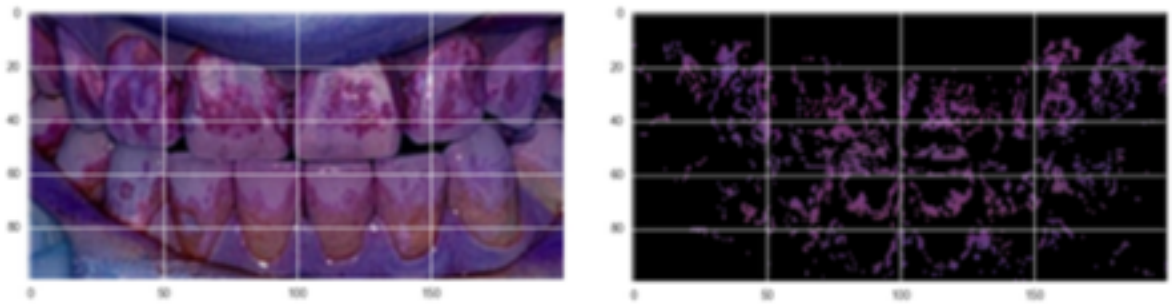


Figure 2: Computer Vision algorithm at play

We also tried feeding all the 20,000 pixels of binary variables into our models. Additionally, we also tried to group the 20,000 pixels into 8 bins - partitioning the image into 8 sections, counting the number of pink pixels in each and adding them as 8 features instead of 1 or 20,000.

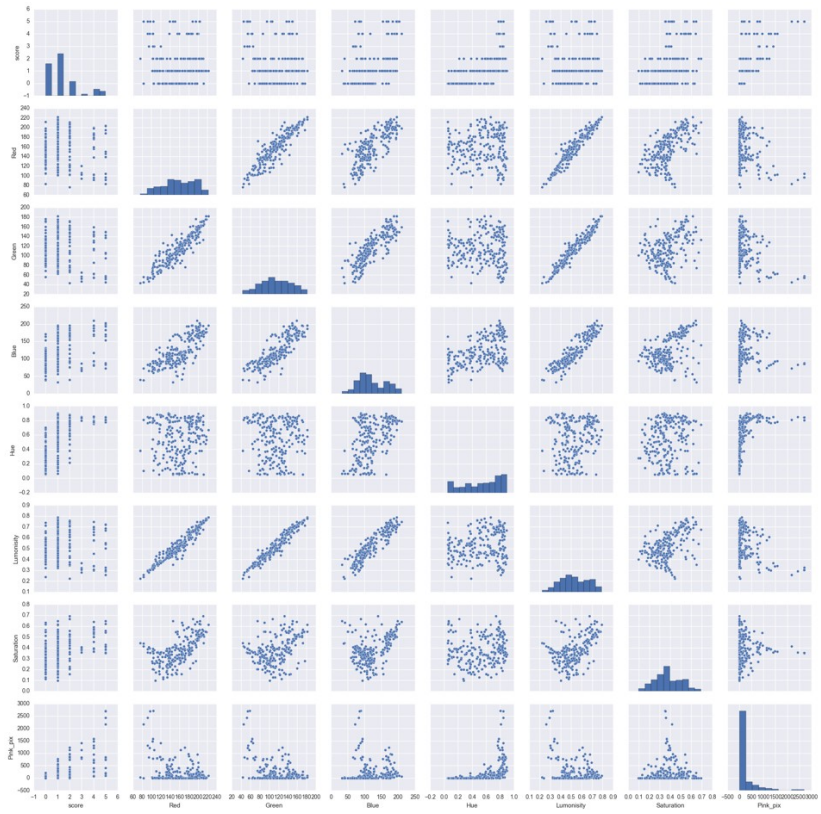


Figure 3: Scatterplot matrix of the target variable and final feature set—total pink pixels count and average RGB and HLS values. The diagonals are the distribution plots (See Appendix B). 'Score' is the target variable.

Since RGB, HLS and pink pixel counts are features on different scales, they have been z-score normalized before being fed into the model. Looking at score (the label) against other features plot, these features are non-linear with respect to score. On top of it, many of these features, especially R,G,B, show an interaction effect because the 'pinkness' which represent plaques depend on all three R,G,B values. Therefore, instead of applying non-linear transformations on each of these features and constructing explicit features to capture interaction effect between them, algorithms such as Random Forest and SVM (Kernel generate transformations implicitly) have been considered to handle these non-linear features with interaction effects.

Modeling & Evaluation

When assessing the performance of models, we used MSE (Mean Squared Error) as an evaluation metric which means that predicted values that are further away from the correct value get penalized more. This is because we are more concerned about informing the user about their plaque condition and a prediction that is 1 off is much better than a prediction that is 5 off. Accuracy is used also as secondary evaluation metric to assess in the case of low MSE whether the model is mostly getting correct values but with few of them being really off (high accuracy) or just close to the true value in general but not many correct values (low accuracy). Additionally, it also offers a more intuitive insight to communicate the predictive power of our model to interested non-technical parties whether they are investors or potential users.

We started with two baseline models. The first is the model that always picks the modal class 1, which is the class that appeared the most in the dataset. The second is the model that always picks the middle class 3 since our evaluation metric is MSE and the error might be minimized this way. The MSE performance for them is 1.80 and 4.90 respectively, whereas the accuracy is 41% and 2% respectively. Since the dataset is heavily skewed with the vast majority being class 1, the modal class has been used as baseline to evaluate other models, but once deployed and new datasets have been obtained using middle

class 3 might be more suitable for MSE performance evaluation.

We considered many data mining algorithms for this project. The first was One Versus Rest (OVR) Logistic Regression since it is fast, scalable, and easy to implement. OVR Logistic Regression out of the box is also a good alternative baseline model to evaluate how other data algorithm perform compared to the simplest one available. Since our target variable which ranges from 0-5 is ordinal, Ordinal Logistic Regression fits better with the ordinal nature of the target variable. Ordinal Logistic Regression algorithm was obtained from the following source: http://pythonhosted.org/mord/, and the algorithm is more limited in terms of parameter options and might not be as sophisticated as Logistic Regression from scikit-learn. However, a common drawback for Logistic Regression algorithms which applies here is that they do not handle non-linear features with interactions effects well.

We also considered SVM because the use of Kernels allows for implicit non-linear transformations of features with interaction effects. SVM with RBF (Radial Basis Function) kernel has been chosen specifically because 'poly' kernel takes too long to run (almost 1 day and not a single result). Due to the radial function being computationally expensive, training SVM with RBF kernel can be slow.

Lastly, we considered Random Forest. It handles high dimensional space and non-linear features that might interact with each other well. Most importantly, it is easy to use off the shelf and tends to perform well due to the wisdom of crowd strategy it utilizes. However, because it has to keep multiple trees in storage the model might be slow to evaluate and be memory consuming.

Since our dataset is small, having only 468 pictures, and is heavily skewed with only 8 class 3 samples, stratified sampling has been used for splitting the dataset into training and testing (as mentioned in the Data Understanding part), and stratified K fold is used for cross validation to ensure that the model learns from and is tested against every class. In each iteration of evaluation, the random state of sampling and splitting has been set to the same value/seed to ensure that splitting is identical (same indices in train set, test set and validation set) for different features sets so that they are comparable and consistent in terms of performance. Multiple evaluations are run to obtain an average value and standard error for the performance of the models. The evaluation framework involves comparing performance of multiple data

algorithms with different set of features, and depending on whether the key feature seems useful further feature engineering is done on the related feature to further improve the performance. For instance, adding in total count of Pink Pixels to the model increased the performance, therefore we decided to add the 8 partitions of pink pixel counts expecting it to further improve the performance. Although the baseline is always picking the modal class 1, the default logistic regression is instead used more often to compare how the other models are doing with the newly introduced features. Each model is tuned using grid search with stratified K fold and scoring = negative MSE to find the parameters that would give the best performance according to the validation set, except for Logistic Regression for which we used the 1 standard error rule. The performance of the default model is also computed to compare how tuning each model affects its performance.

During our first preliminary results, we started out with 4 different feature sets to test the performance of different image preprocessing: 1. original image with 6 features which consist of only RGB and HLS average values; 2. original image (20000 pixels) with 60000 features each of which is either R,G or B values; 3. gray-scaled image containing 20000 features each of which is a pixel value of RGB (R,G,B have the exact same value when gray-scaled) ; 4. blurred image containing 60000 features each of which is RGB but each averaged around with the neighbor's values. Based on the best performing algorithm (detailed comparison of algorithms performance is discussed later on), tuned random forest which gave the best results across all the feature sets, the best performing model was the feature set 1 with a MSE of about 0.15. (The Logistic Regression Model Performance for each feature set is located in Appendix C).

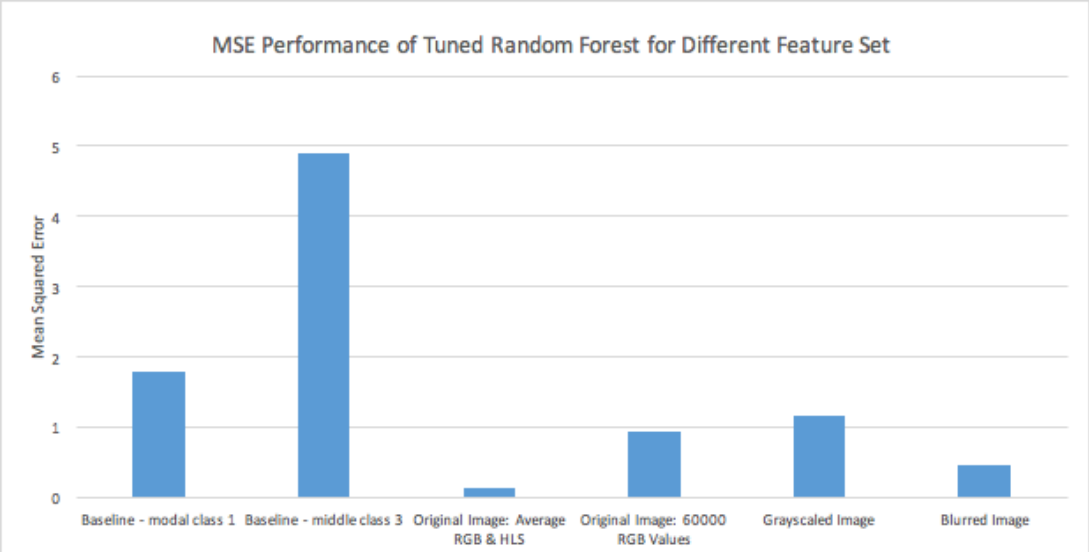


Figure 4: MSE performance of the best performing algorithm across all the feature sets

It is possible that the original image with RGB values and grayscale images are noisy because each pixel represents part of a mouth but every image is displaced differently. Since the center of an image is not uniform across all the picture, the features that represent each pixel could be learning this displacement noise instead. The model with blurred image performs better than those two, perhaps because it has averaged the neighboring values, which might have reduced noise but at the same time quality of 'pinkness' detection of plaque. Thus, in the evaluation process we decided to leave out image preprocessing techniques, such as grey-scaling and blurring, after seeing decrease in performance using these features.

Based on the fact that average RGB and HLS gave the best results, we speculated that it could be due to how average RGB and HLS were good proxies for pinkness of the image, which represent level of plaque. Therefore, to further improve our model we focused on Computer Vision to extract features relevant to detecting the pinkness of the image. As illustrated in images from 'Data Preparation', new features related to 'pink' which represent plaque amount has been introduced into the original feature set that has only average RGB and HLS values. In this new iteration of evaluation, we had the following

feature sets: 1. original feature set which was the best performing one from the first preliminary results; 2. original feature set with an additional feature, the total count of pink pixels; 3. extending the idea from 2. we divide the total pink pixel count into 8 different partitions each one with its own count of pink pixels plus the original feature; 4. the union of the 3 feature sets mentioned above; 5. original feature set and 20000 features each of which is a binary value representing whether the pixel is 'pink'. Note that this feature set 5. was left out at an early stage because compared to the other feature set it gave a poor MSE performance of 1.3191 by the tuned random forest and 1.1064 by the tuned Logistic Regression and it was taking considerably long time to run (see appendix D for the performance plot for this feature set). During this process, as shown in the figure below we have noticed significant discrepancy between the performance of Logistic Regression and Random Forest which is likely due to the non-linear nature of features which also have interaction effect (see Data Preparation for details). Therefore, SVM with RBF (Radial Basis Function) kernel has been included for performance evaluation so that Random Forest is not the only obvious candidate since Logistic Regression cannot handle that.

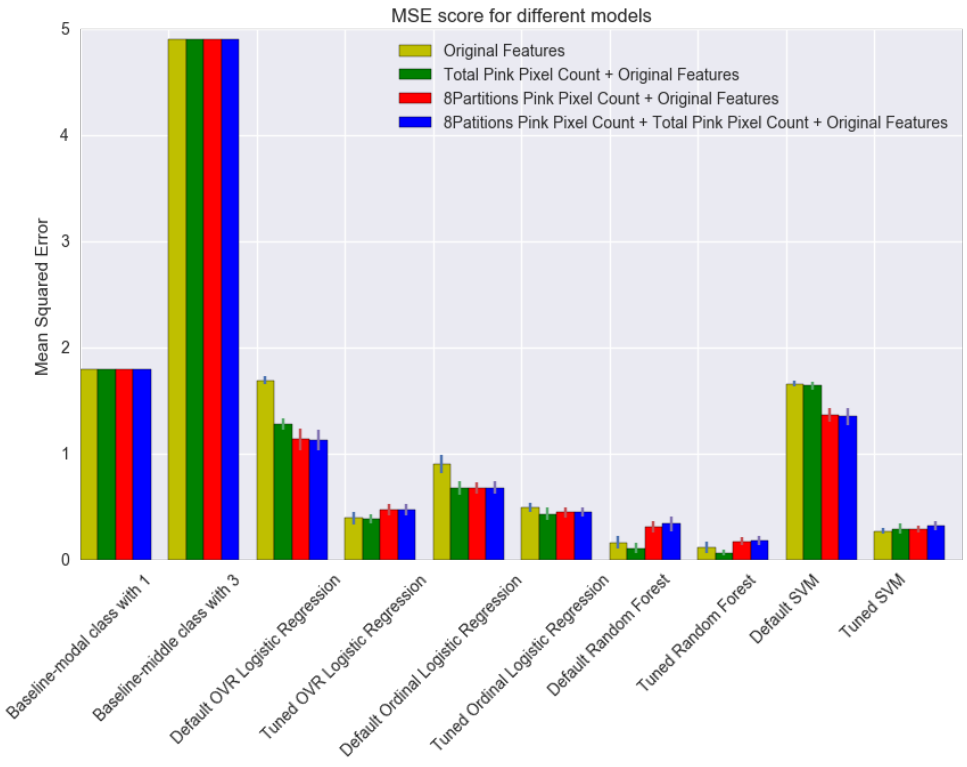


Figure 5: MSE scores for the different models

According to the final evaluation shown in the figure above, the best performing model is Tuned Random Forest with feature set 2 - the total count of pixels, RGB and HLS average values. As expected, both Logistic Regressions models, OVR (One Versus Rest) and Ordinal, performed worse across all features due to non-linear features with interactions effect present in the dataset. SVM with tuned parameters performed quite well in general compared to Logistic Regression, but not as optimal as Random Forest. Tuning parameters also have had significant impact on the performance of algorithms in general, where the MSE performance of the best performing algorithm, Random Forest, went from 0.11 down to 0.07 and 93% up to 95% in terms of accuracy (See Appendix E for the parameters of the best model, Tuned Random Forest with feature set 2. See Appendix F for accuracy scores).

Surprisingly, although the feature set 3, which consist of original feature set plus 8 partitions of pixel counts, is an extension to feature set 2, which has simply total count instead of 8 partitions of pixel, models with feature set 3 performed worse than models with feature set 2 in almost all the algorithms. Similarly, models with feature set 4 which is the union of feature set 2 and 3, also performed worse than models with feature set 2. A possible explanation is that by making pink pixels count position/partition dependent, more noise might have been introduced because images are not taken in a uniform way, just like how during the first preliminary results where models with 60000 features of RGB values were performing worse than simple average of RGB and HLS values. Hence, we conclude that Random Forest with feature set 2 and tuned parameters is the best performing algorithm with an outstanding performance of 0.07 MSE and an accuracy of 95%, a significant improvement over the baseline models. In fact, looking at the confusion matrix of this model, we see that the model makes most of predictions correctly, and the few incorrect ones are only off by 1. However, as observed in the confusion matrix, the distribution is so skewed towards class 1 that misclassification often happens around class 0 and 2, where they get incorrectly classified as 1 instead.

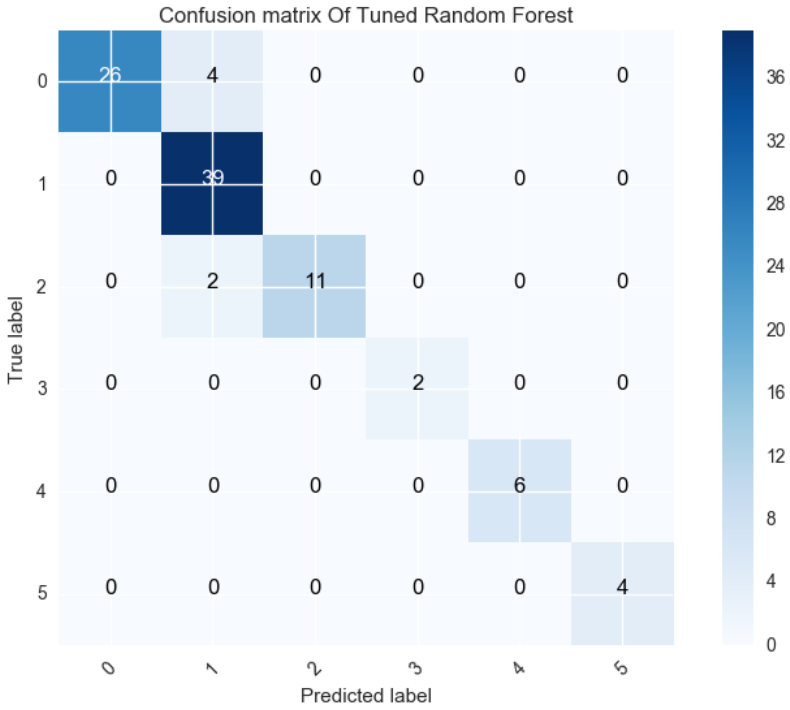


Figure 6: Confusion matrix of tuned random forest

In consideration of using tuned Random Forest with RGB, HLS average values and total pink pixel count as the model for deployment, we inspected the learning curve and the validation curve of the model to seek for possible room of improvements.

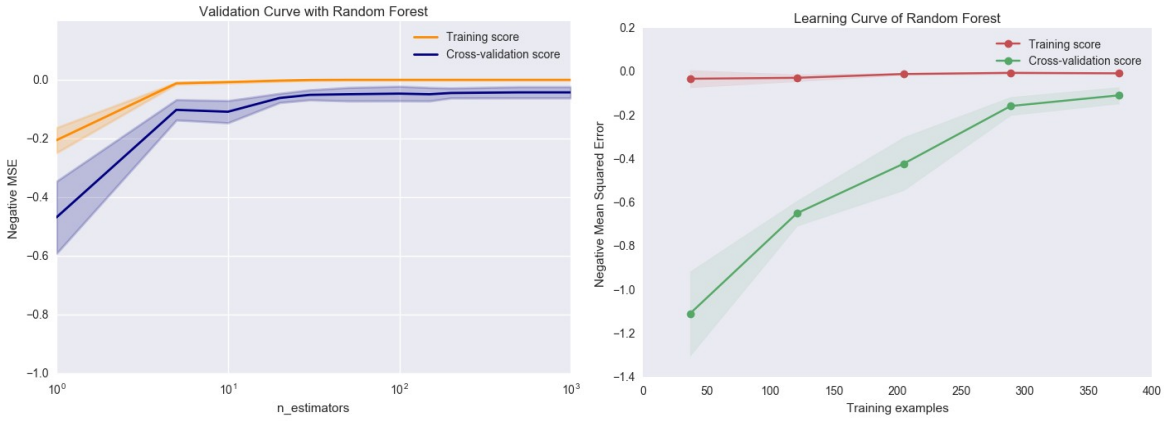


Figure 7 (Left): The validation curve plot of the best performing model, tuned random forest with feature set 2

Figure 8: (Right): The learning curve plot of the same model

As shown in validation curve plot, the model achieves best performance when number of estimators is 100, but it achieves roughly the same performance around 50. Therefore, number of estimators for this model is picked as 50 once deployed, because that might effectively reduce the running cost of the model by half by cutting down the number of estimators by a half.

Looking at the learning curve, we also see that the performance of the model albeit minimal might further improve with more training samples. However, once we start obtaining data from a deployed environment, the performance of the model will most likely go down. As a matter of fact, a large factor in which the model performed so well, especially compared to baseline model, is because the images obtained from dental students are well-structured and clean. In fact, all images are exactly 20000 pixels and there are no visible blurs in any of the images, which is difficult to expect in a deployed environment. Therefore, as part of future areas of improvement, it'd be helpful to include pictures that have been taken by potential users. In addition to that, Singular Value Decomposition might be used as feature extraction to separate the features so that they are linearly independent, in which case one of the vectors which represent noise could be removed (if it exists) in order to improve the performance of models, especially the ones with feature set 3, 8 partitions of pixel count.

Our model offers a prediction of plaque level that is very close to its true value, providing a lot of value to its users in terms of dental care and therefore opens doors for many business opportunities which will be expanded upon in the deployment section.

Deployment

This project can be deployed as a phone app. Since the disclosing tablets cost a few pennies, the app is an accessible way to track users' oral health. Thus, we believe business plan with emphasis on empowerment via education and personal healthcare will be the most successful. Despite being preventable, cavities are a widespread and growing problem for large subpopulations of America (See Appendix A). Our app is a simple and cheap way to promote the long-term oral self-care needed to

prevent the disease.

A large part of cavity prevention and dental plaque management is the continuous maintenance of oral health. Due to the infrequency of dental appointments—in 2014, 33.3% of Americans did not go to the dentist in the past year—it is easy to slip into poor habits and neglect. As part of our marketing plan, we hope to sponsor events at public schools to visually motivate students to recognize and maintain their oral health. On top of the social benefits of this, our project can also be eligible for public funding and tax breaks as government-sponsored events.

Additionally, the app can serve as an advertising and e-commerce platform for dental practices. When a client's plaque level rises above a certain threshold and does not resolve, the app can recommend a visit to a dentist and list local practices based off of the user's current position. To motivate the client, the app offers convenience in making appointments and alerting the dentist. Monetization is modeled after the Seamless/GrubHub platform, where dentists pay a percent royalty of the services booked through the app. Higher percentages will allow certain practices to be promoted higher on the list of local services. Additionally, the company can add 24/7 customer support, to ensure that scheduling and payment conflicts can be resolved to protect the client.

The ethical issues with deployment are tied with the Type I and II errors from the actual Machine Learning algorithm. In the Type I Error/False Positive case, where the app classifies the client to a higher amount of plaque than the true amount, a client may visit a dental practice despite maintaining healthy plaque levels. An ethical conundrum lies in the monetization of advertising medical practices. From an amoral perspective, both the company and dental practices will profit greatly from collusion by abusing false positives. Thus, it is critical from an ethical perspective to ensure that every practice on the app is vetted as a licensed dental service provider, and a system in place to prevent abuse.

The Type II Error/False Negative case, where users are classified with a lower score than the actual amount, presents a very different ethical issue. A user can garner a false sense of security if we predict a low score (good teeth) when they actually have a high score (bad teeth)—as a consequence, the users' dental health will be compromised and the app will be open to litigation. The Machine Learning

algorithm at hand works well with a dataset that was collected at one hackathon by a single group of individuals. In practice, the images that clients upload may have issues of resolution and lighting. A final consideration lies in an initial filtering system to see if the image is actually an image of teeth, so that the clients do not lose faith in the models' predictive power.

In the future, the app can use Computer Vision boundaries to recommend brushing certain areas or to use certain techniques, such as the Modified Bass technique for brushing. Informative videos and diagrams can be included, to emphasize the focus on education and continuous cleaning. Recent advancement disclosing tablets contain multiple dyes to stain plaque different colors corresponding to age rather than uniformly staining plaque pink.[4] Therefore, this project not only has the potential to provide a lot of value with current technology, but also has much room for future expansion to provide a more holistic dental diagnostic.

[4] D.J. Fischer et al. *Dental Caries and Periodontal Conditions, in Risk Assessment and Oral Diagnostics in Clinical Dentistry*., Chapter 6.

Reference List

Broadbent, Jonathan M., W. Murray Thomson, John V. Boyens, and Richie Poulton. "Dental Plaque and Oral Health during the First 32 Years of Life." The Journal of the American Dental Association 142, no. 4 (April 2011): 415-26. doi:10.14219/jada.archive.2011.0197.

"Dental Caries (Tooth Decay)." National Institutes of Health. May 28, 2014. https://www.nidcr.nih.gov/datastatistics/finddatabytopic/dentalcaries/.

Fischer, D. J., Treister, N. S. and Pinto, A. (2013) Dental Caries and Periodontal Conditions, in Risk Assessment and Oral Diagnostics in Clinical Dentistry., John Wiley & Sons, Inc., West Sussex, UK. doi: 10.1002/9781118783283.ch6

Kapner, Michael, DDS. "Dental Plaque Identification at Home." University of Maryland Medical Center. http://umm.edu/health/medical/ency/articles/dental-plaque-identification-at-home.

Peterson, Scott N., Erik Snesrud, Jia Liu, Ana C. Ong, Mogens Kilian, Nicholas J. Schork, and Walter Bretz. "The Dental Plaque Microbiome in Health and Disease." PLoS ONE 8, no. 3 (March 08, 2013). doi:10.1371/journal.pone.0058487.

Sipek, Sarah. 2016. "Smile Wide for the Camera." Workforce 95, no. 7: 15. Business Source Complete, EBSCOhost (accessed December 1, 2016).

United States of America. Centers for Disease Control (CDC). National Center for Health Statistics. “Health, United States, 2015”. By Sylvia M. Burwell, Thomas R. Frieden, M.D., M.P.H., and Charles S. Rothwell, M.S., M.B.A. U.S. Department of Health and Human Services, 2016.

Appendix A

A Brief Overview of Dental Caries

Dental caries is a product of time, certain strains of bacteria, and carbohydrates on the surface of a tooth. The biofilm that surrounds teeth, dental plaque, is comprised of a whole community of microorganisms. Some strains of bacteria, when allowed to survive, will take in carbohydrates and multiply. Acid is a byproduct of this interaction, which dissolves the enamel surface of teeth and causes a cavity.[5]

Prevention measures mostly revolve around maintaining good oral self-care habits to periodically clean the surfaces of one's teeth and remove dental plaque. Despite this, the percentages of Americans with untreated caries in on the rise, with 27.1% of Americans aged 20-44 years and 25.8% of Americans aged 45-64 years have been found with untreated tooth decay from 2011 to 2012. By contrast, the same age demographics posted 25.1% and 21.6% rates of untreated dental caries respectively in the time span between 2005 and 2008.[6]

Improving oral health requires long-term self-care and openly accessible public health measures that educate individuals on proper dental hygiene.[7] This includes actual techniques like flossing or the Modified Bass method of brushing, but also an emphasis on continuous dental self-care.

---

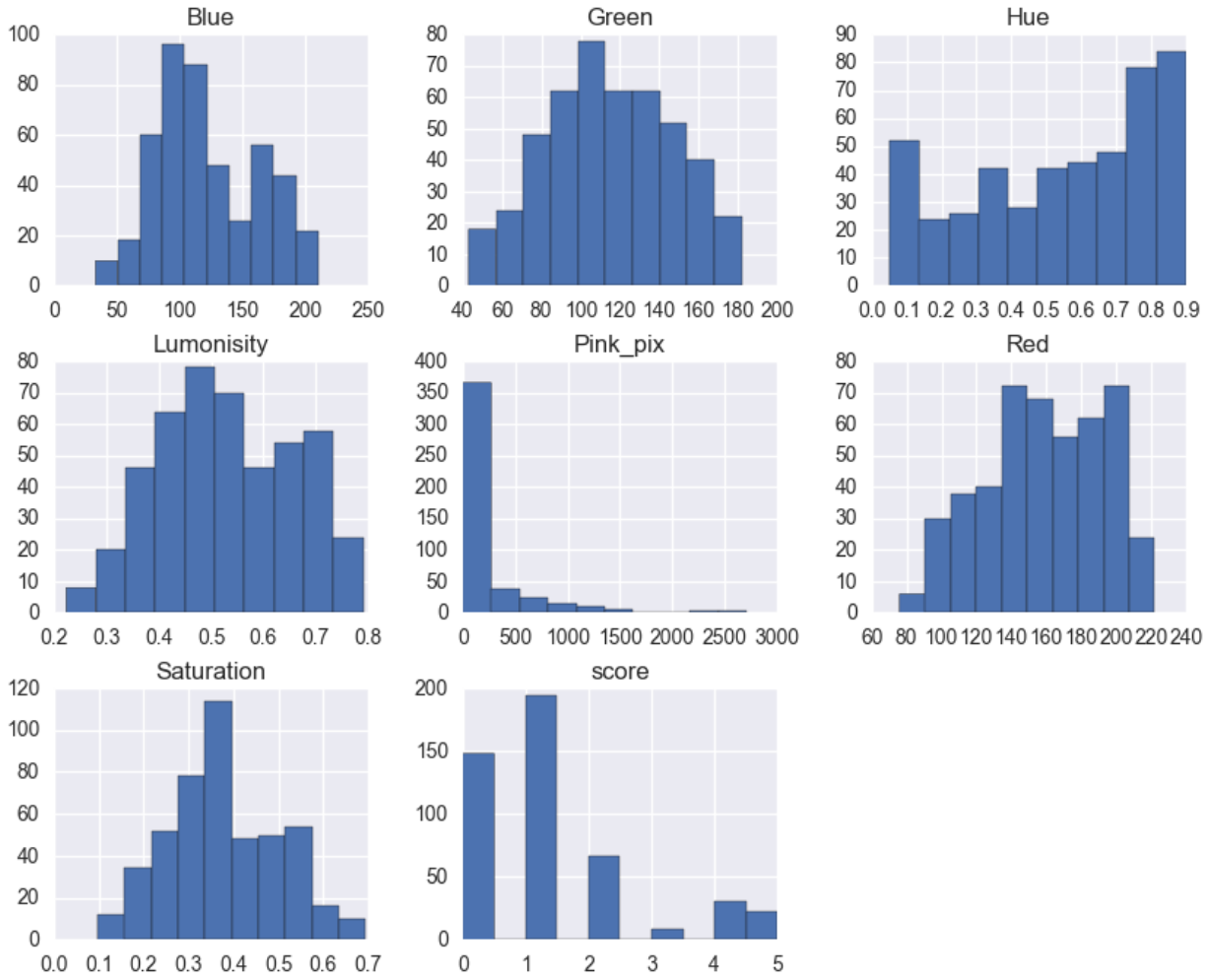
[5] D.J. Fischer et al. *Dental Caries and Periodontal Conditions, in Risk Assessment and Oral Diagnostics in Clinical Dentistry*., Chapter 6.

[6] Centers for Disease Control (CDC)., "Health, United States, 2015".

[7] Jonathan Broadbent et al., "Dental Plaque and Oral Health during the First 32 Years of Life."

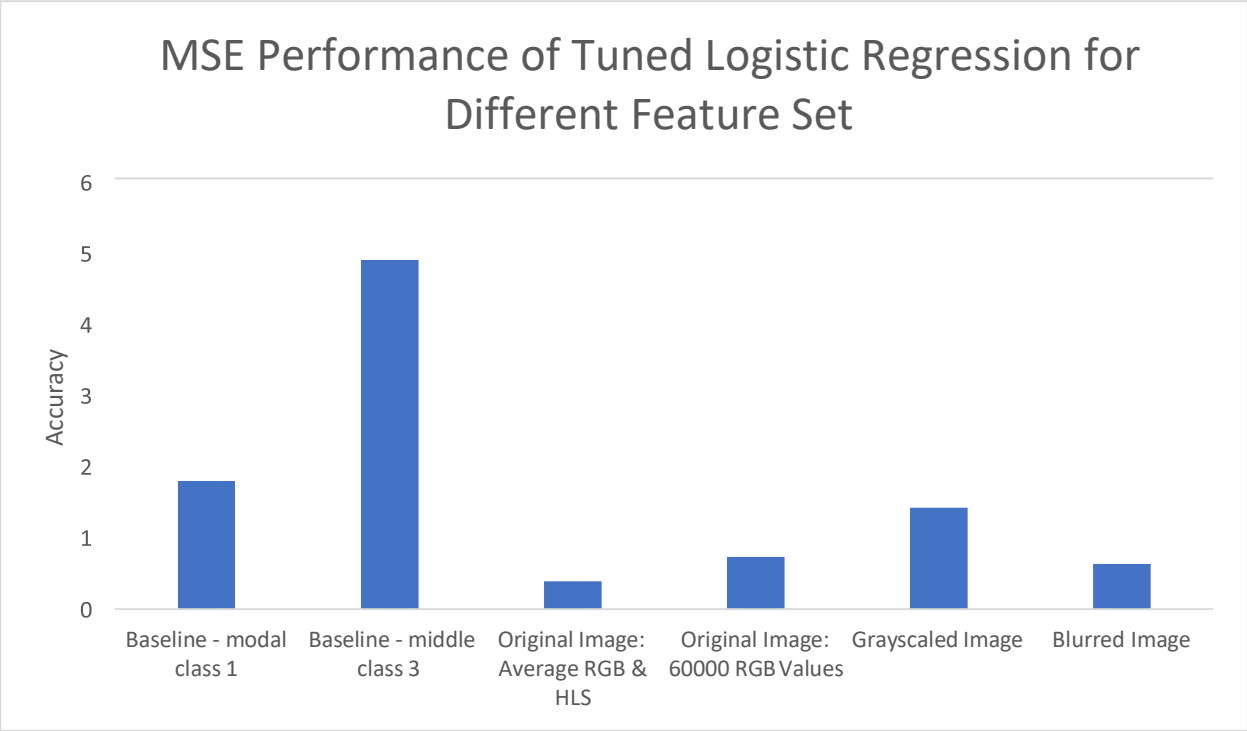
Appendix B

The distribution graph of the final feature set – RGB & HLS average values and pink pixel count
Last graph is the distribution of the target variable, score.



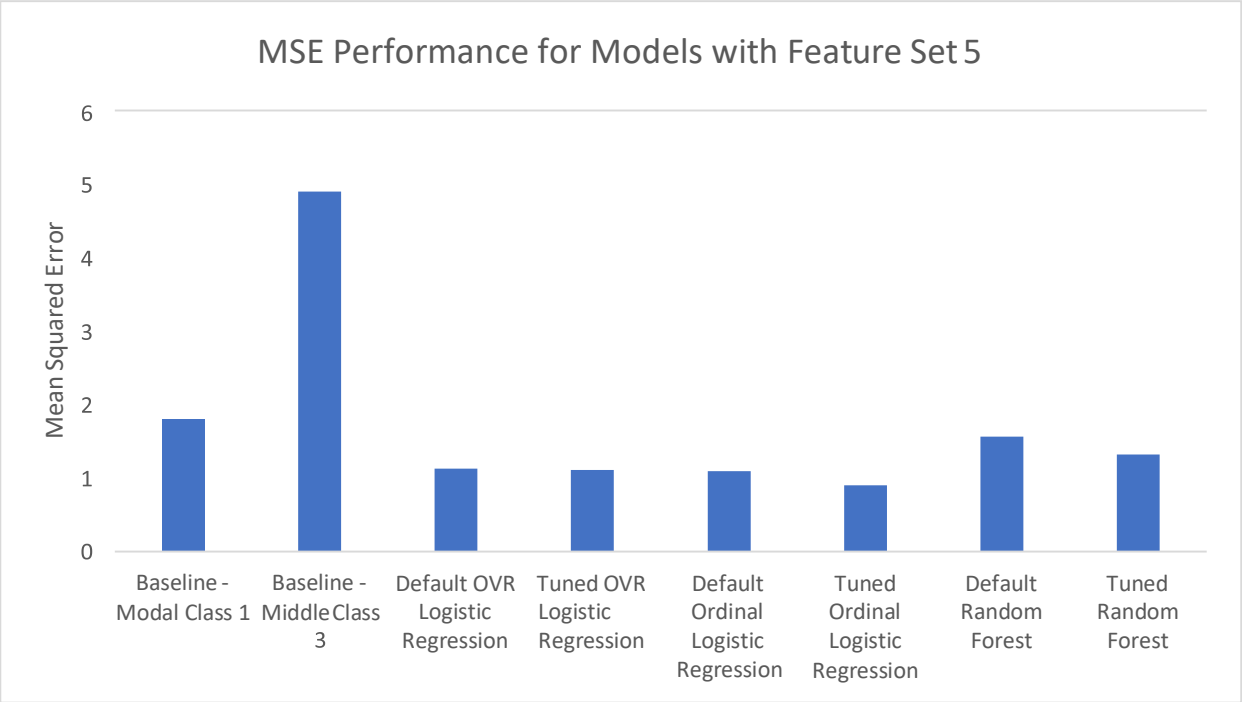
Appendix C

MSE performance of OVR Logistic Regression across all the feature sets proposed during the first round of preliminary results



Appendix D

The mean squared error for models with feature set 5 – RGB & HLS average values and 20000 features each of which is a binary value representing whether the pixel is 'pink'



Surprisingly, Random Forest with tuned parameters performed worse than the Logistic Regression.

Appendix E

The tuned parameters for Random Forest, the best performing algorithm, with feature set 2 (total pink pixel count + RGB&HLS average values)

Parameters of the Tuned Random Forest:

{'warm_start': False, 'oob_score': False, 'n_jobs': -1, 'verbose': 0, 'max_leaf_nodes': None, 'bootstrap': True, 'min_samples_leaf': 1, 'n_estimators': 100, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'criterion': 'gini', 'random_state': 1, 'min_impurity_split': 1e-07, 'max_features': 'auto', 'max_depth': None, 'class_weight': None}

After checking the validation curve, we decided on using n_estimators = 50 to retain about the same performance but at half of computational cost.

Appendix F

Average MSE performance for each model. The error bars represent 95% confidence interval. Yellow represent feature set 1 with only RGB and HLS values; green represent feature set 2; red represent feature set 3; the blue represent the union of all the features mentioned above (because it was observed that the feature set with only total pink count to perform the best so we expected blue would perform even better for including both the 8 partitions and the total count).