Machine Learning Classifiers on Traditional and Modern Tobacco Use

Application Track, Jiangyue Mao, Brandon Apodaca, Jiale Tan

Abstract

Tobacco use has an evidence-based association with many chronic diseases such as cancer, cardiovascular disease, mental disorders, etc. Although combustible cigarette use has declined in the US, the percentage of some users of modern tobacco products such as e-cigarette users who initiated at a young age is 3 times higher in 2018 than 2014 [3]. Therefore, more attention should be focused on modern tobacco products with increasing prevalence. However, most machine learning papers only focus on traditional combustible cigarette use. This motivated us to use machine learning methods to classify an individuals use status of modern tobacco products (Never user, Former user, and Current user) based on their use of other tobacco products combined with sociodemographic factors including age, ethnicity, and marital status. Our results show that the Kernel Ridge Regression yields the best performance when compared with other machine learning methods when evaluated using MSE and AUC scores. Future work includes the application of these machine learning methods to a more balanced dataset and the inclusion of more information-rich features.

1 Introduction and Motivation

Tobacco use is associated with a high risk of contracting many diseases such as cancer, cardiovascular disease, and pulmonary disease, all of which can result in death. It is also a risk factor for respiratory tract disease and other infections, osteoporosis, reproductive disorders, adverse postoperative events, delayed wound healing, duodenal and gastric ulcers, and diabetes [1]. Although there are several common carcinogens in tobacco products, different tobacco products are often associated with different health risks [2]. Harmful metals such as lead, cadmium, and beryllium produced by incomplete combustion of traditional cigarettes are not present in the same concentrations in other tobacco products. For example, some E-cigarettes produce low metal concentrations in comparison to cigarettes since they do not burn tobacco. Even though combustible cigarette use has declined in the US, the percentage of lifetime e-cigarette users who initiated use at 14 years or younger increased from 8.8% in 2014 to 28.6 % in 2018 [3]. This increase use of modern tobacco products such as E-cigarettes and others arouses a great public health concern.

We divide tobacco product use status into three groups: Never use, Former use, and Current use. Never use means participants have never smoked before, former use represents smokers who have not smoked in the past 30 days, and current use represents those who have used tobacco products at least once in the past 30 days. We aim to classify a user's modern tobacco use status based on their use of other tobacco products combined with sociodemographic factors such as age, race, marital status etc.

2 Problem Statement

In this project, we will use six machine learning methods to classify tobacco use status using an individual's use status of other tobacco products combined with sociodemographic factors. The conceptual and math notations are displayed in Table 1. While relevant mathematical equations and variables for all models used are shown in Table 1, the external library, sklearn, was imported and used for actual implementation due to its prior optimization with respect to the python coding language in addition to the availability of built-in evaluation functions.

Table 1: Model Equations

Model	Parameters	Objective Function
Kernelized SVM	α_i, y_i: label for sample i $\kappa(x_i, x_j) = \psi(x_i)\psi(x_j)$	$\max_{\alpha \in R^n} \Big\{ \sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \Big\}$ s.t. 1) $0 \leq \alpha_j \leq C$ 2) $\sum_{j=1}^{n} \alpha_j y_j = 0$
Kernel Ridge Regression	α, y: label vector i K : Kernel Matrix λ : Regularization Parameters	$\min_{\alpha \in R^n} \frac{1}{2} \|\|y - K\alpha\|\|_2^2 + \frac{\lambda}{2} \alpha^T K \alpha$
Logistic Regression	σ: logistic sigmoid function w: feature weights y_i: label for sample i	$\min_{w} \sum_{j=1}^{n} (\sigma(w_0 + w_1^T x_j) - y_j)^2$ $\sigma(u) = \frac{1}{1 + \exp(-u)}$
Lasso Linear Regression	w: feature weights λ: regularization parameters y_i: label of sample i	$\min_{w} \sum_{j=1}^{n} \big(y_i - w^T x_j\big)^2 + \lambda \|\|w\|\|$
Random Forest	pP: proportion of class 1 in parent node pL: proportion of class 1 in left child pR: proportion of class 1 in right child q: proportion of data in parent node that goes to left child H_p: $p \cdot log_2 p - (1-p) \cdot log_2(1-p)$	$IG = H(pP) - [qH(pL) + (1-q)H(pR)]$
Multilayer Perceptron	$(e_j)^2$: $y_j - y_j(v_j)$ $y_j(v_j)$: $\frac{1}{(1+exp(-v_j))}$ y_j: label of sample j	$\min_{vi} \frac{1}{2} \sum_{j=1} (e_j)^2$

3 Related Work

There are currently four types of machine learning applications in tobacco research: (1) Machine learning-powered technology to assist traditional cigarette cessation [4,5,6]. (2) Content analysis of tobacco on social media [7,8,9,10]. (3) Tobacco-related outcome prediction using administrative, survey, or clinical trial data [11]. (4) Smoker status classification from narrative clinical trials [12,13]. Based on this we note that almost all machine learning tobacco research is about traditional cigarettes. Other modern and minority tobacco products such as cigarillo, hookah, smokeless, pipe, dissolvable tobacco etc. have yet to be thoroughly analyzed using machine learning algorithms. The corresponding details are shown in Appendix Table 5.

4 Methodology

4.1 Dataset

The Population Assessment of Tobacco and Health (PATH) Study in the United States came from the Inter-University Consortium for Political and Social Research (ICPSR) and is available for public use. The database contains multiple datasets, each with different variables and around 50,000 responses from just the third wave of the surveys. Variables in different datasets contain different information, ranging from the frequency usage of many kinds of tobaccos such as cigar, hookah, and smokeless tobaccos etc. to demographic information such as educational level, race, age, medical conditions, marital status etc. All the non-demographic features are binary, using 1 to represent True and 2 to represent False to indicate the medical and smoking status of the patients.

4.2 Data Preparation

In order to perform data processing and analysis, we first clean the data. According to the code book provided alongside the data, all negative values represent instances when a subject declined to answer the question or data is missing. The existence of a NaN value also represents an incomplete sample. Therefore, all samples containing at least one negative or NaN entry must be eliminated to yield a complete dataset.

The original variables directly from the dataset do not contain the desired classification information for this project, we must produce the ground truth labels for each sample in preparation for the machine learning task. The variables directly from the dataset include the following for each tobacco product: "R03R_A_P30D_[TOBACCO_PRODUCT]" and "R03M_EVR_[TOBACCO_PRODUCT]" where [TOBACCO_PRODUCT] is replaced by the abbreviated version of each tobacco product name. For example, for cigarettes, these variables are, "R03R_A_P30D_CIGS" and "R03M_EVR_CIGS". Assuming all individuals are telling the truth, the first variable represents whether or not that individual has used a given tobacco product in the last 30 days. The second variable represents whether or not that individual has ever used the given tobacco product. These abbreviations are shown in Table 2. Logic statements can be used to generate labels which represent current users, former users, and individuals who have never used a given tobacco product. If an individual has not used a given tobacco product in the last 30 days but has used that tobacco product before, then they are classified as a 'former user' of that tobacco product. If an individual has never used a given tobacco product before, they are classified as a 'never user' of that tobacco product. Otherwise, the individual is classified as a 'current user' of that tobacco product. This logic is applied to every tobacco product and appended to the demographic variables to form the complete dataset. To be specific, for each tobacco product, we create three new boolean variables for current users, former users, and never users, where only one of which will be true for a given individual.

Table 2: Tobacco Product Abbreviations

Tobacco Product	Abbreviation
Cigarettes	CIGS
E-Product	EPRODS
Filtered Cigar	GFILTR
Cigarillo	GRILLO
Traditional Cigar	GTRAD
Hookah	HOOK
Tobacco Pipe	PIPE
Smokeless Tobacco	SMKLS
SNUS Pouch	SNUS

For each tobacco product, the data is segmented into training features and ground truth labels. The three binary labels ('former user', 'current user', and 'never user') represent the ground truth labels for a given tobacco product ($\mathbf{y}$). The remaining features including demographic data and user status for the other tobacco products are used as the training features ($\mathbf{X}$). The resulting data set, $\mathbf{S} = \{\mathbf{X}, \mathbf{y}\}$, is split into a training testing set (75% of S) and a testing set (25% of S). For example, classification of a current user of a smokeless tobacco product requires the use of the label and feature combination shown in Table 3

Table 3: Example of Features and Label Combination for a current user of Smokeless tobacco products

Label	Features
SMKLS_current	CIGS_current, CIGS_former, CIGS_never, EPRODS_current, EPRODS_former, EPRODS_never, GFILTR_current, GFILTR_former, GFILTR_never, GRILLO_current, GRILLO_former, GRILLO_never, GTRAD_current, GTRAD_former, GTRAD_never, HOOK_current, HOOK_former, HOOK_never, PIPE_current, PIPE_former, PIPE_never, SMKLS_former, SMKLS_never, SNUS_current, SNUS_former, SNUS_never, Marital_single, Marital_married, Marital_divorced, Age, Race

4.3 Classification Methods

The classification task was accomplished by applying each model to each dataset for each tobacco product. Since there are three binary classes per tobacco product, three models of each type were trained for each product. The models used to perform classification are listed in Table 1. All models used except the decision tree are kernelized. This means model weights do not correspond to specific features, rather they correspond to the features resulting from the kernel function.

In order to perform model selection, a variety of models were applied to each dataset. Both classification and regression methods were used, in addition to a decision tree. For each model, a grid search using cross validation was applied to select the hyperparameters which yielded the lowest mean squared error.

All dataset processing and machine learning algorithms were implemented in python. Dataset preprocessing was implemented using the pandas and numpy libraries. Machine learning models and grid search using cross validation were implemented using the sklearn library. The sklearn library simply implements machine learning models by providing a model initialization function to produce a model object. After initialization, train and predict member functions for each model object can be called to apply data to the training process and generate classification probabilities for all test samples.

5 Evaluation

Since our target is to predict the type of tobacco user(i.e. ever user, current user, or never user), we apply evaluation metrics to each type of tobacco user to observe the interpretability and predictability of each model. Evaluation metrics include Akaike information criterion(AIC), Bayesian information criterion(BIC), Area Under Curve(AUC), and mean squared error (MSE). AIC and BIC are useful for interpretability. They reveal the degree to which the model fits the training dataset. AUC and MSE indicate how well the model can predict the type of users, a metric of predictability.

AIC can be calculated using the formula $2k - 2ln(\hat{L})$, where k is the number of model features and L is the maximum value of the likelihood function for the model. This Log-likelihood is a measure of model fit. Higher AIC indicates a better fit. In order to simplify the evaluation process, we formulate a uniform AIC formula by assuming the residuals are distributed according to independent identical normal distributions (with zero mean) for each model. Our dataset is relatively large and each patient's response is independent of each other, so making this assumption is consistent with the context of the dataset. Making this assumption gives rise to least squares model fitting. Thus, the maximum likelihood estimate for the variance of a model's residuals distributions is the reduced chi-squared statistic, $\hat{\sigma}^2 = \mathrm{RSS}/n$. And because only differences in AIC are meaningful, some constant terms can be ignored, which gives us AIC = 2k + nln(RSS), where n is the number of samples to amplify the difference. Therefore, AIC represents goodness of fit (as assessed by the RSS), but it also includes a penalty that is proportional to the number of fitted features in the model. This penalty discourages overfitting, which is desired because increasing the number of variables in the model can always improve the goodness of fit. Thus, a lower AIC is associated with a better model fit on the training dateset.

The formula for AIC is similar to BIC, $kln(n) - 2ln(\hat{L})$, where k represents the number of variables, L represents the maximum likelihood estimate of the model, and n represents the number of samples. Here we also use RSS in place of the maximum log-likelihood to simplify the process. We can see that BIC penalizes number of parameters more strongly than AIC, which leads to it becoming a better method to detect overfitting. In this paper, we use both AIC and BIC to examine the goodness of fit.

The True Positive Rate (TPR) represents the diseased patients who received positive results on the test and the False Positive Rate (FPR) represents the proportion of true negatives which are misclassified as positives. The receiver operating characteristic (ROC) curve is the the TPR plotted against the FPR. If the model is perfect at detecting true positives, then FPR will always be 0 and TPR will always be 1. This will produce a curve that is upper triangular and the area under the curve (AUC) will be equal to 1, which is the ideal case. Typically the AUC is less than one, so if the AUC is close to 1, the model is very good at discrimination.

In order to avoid redundant metrics, we did not use accuracy, precision, and f1 score. These metrics are partially redundant with the AUC metric because they are also based on the comparison of true and false positives and negatives.

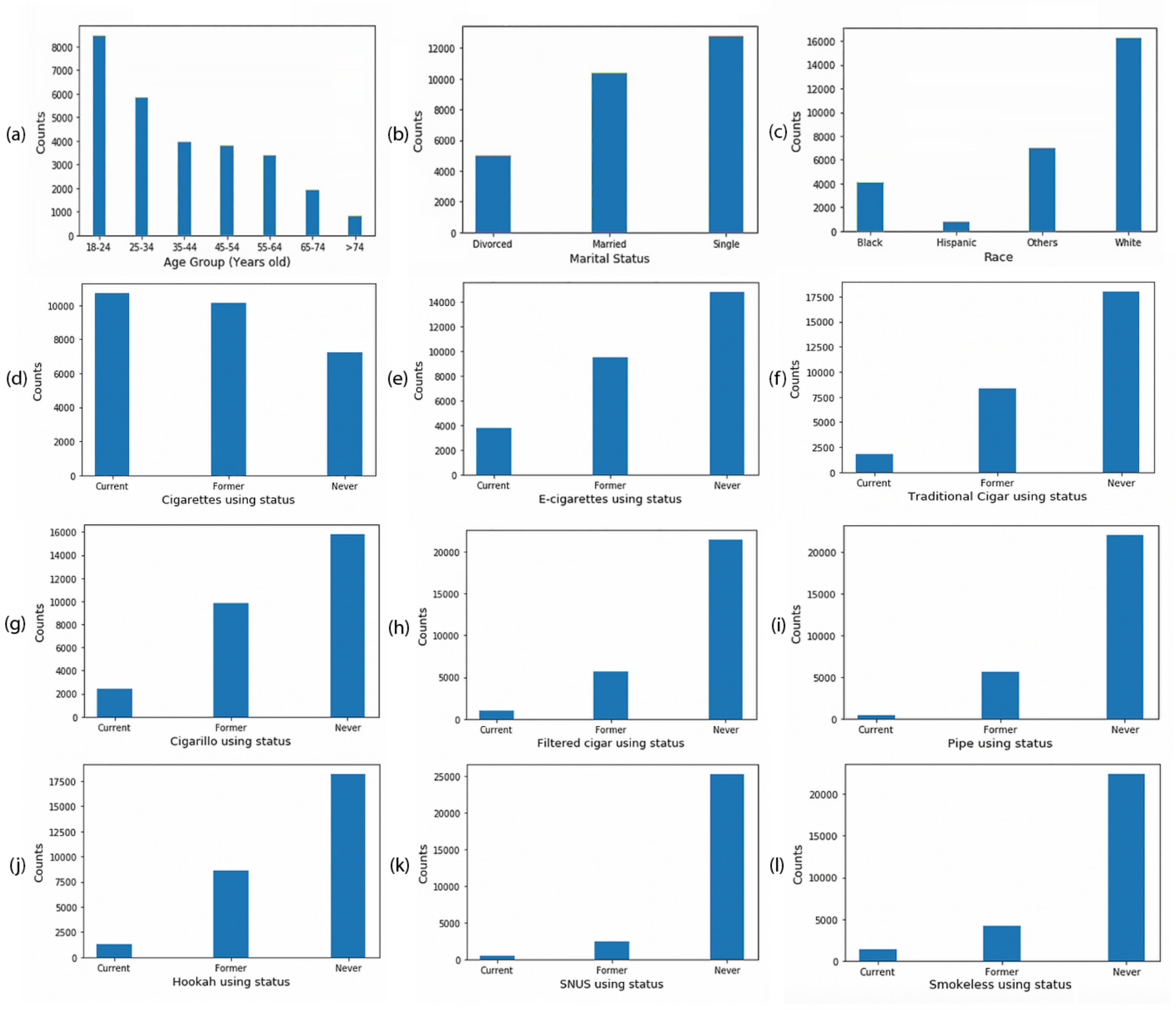
As MSE was mentioned in class, we use the above five metrics to measure the performance of the models(table 1).

6 Results

6.1 Descriptive Analysis

Before implementation of any machine learning models, the empirical data distribution for each variable needs to be considered. Tobacco use status distribution (Never use, Former use, and Current use) of different products, as well as sociodemographic factors such as age, marital status, and race is included in Figure 1. We do not include medical condition and educational status variables because most patients do not wish to disclose this information and only less than 10% of the sample provided this.

Figure 1: Empirical distributions for each feature



By inspection of Figure 1, it is clear that young, single, and white identifying individuals dominated our dataset. In addition, the ratio of current vs never among cigarettes has highest value comparing with other products, indicating participants are more likely to use combustible cigarettes than other modern tobacco products. Among the other, more modern tobacco products, e-cigarettes have highest consumption rate, and Pipe and SNUS have the lowest consumption rate. It is likely that the large imbalance of classes will affect the mean squared error and generate a bias towards low mean squared error values when the classification of never users is dominant.

6.2 Model Fitting Results

The mean of the AIC and BIC scores for all models and every case of user and product combination are 44.4 and 283.1 with standard deviations of 1.5 and 0.8, respectively. This shows that the models have a similar interpretability in terms of every product and type of user. The large BIC score, in comparison to the AIC score indicates that these models are overfitted. This is expected due to the large number of variables in our dataset (32 variables and labels).

Table 6 in the appendix shows that the Support Vector Machine and the Random Forest usually have the lowest MSE, around 0.2 to 0.3, meaning that the difference between the class probability and true label is small on average. The Kernel Ridge Regression has a very good predictability for true positives, yielding the highest AUC value (Table 7), around 0.8 to 0.9 over most cases. The Kernel Ridge Regression also has a similar MSE when compared with other models, within 0.1 of that of the SVM and Random Forest.

We value predictability more than interpretability since our goal is to predict user status. And because the AUC is an estimate of the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, AUC is a better evaluation metric in terms of predictability than classification error rate based upon a single prior probability. Therefore, we choose the Kernel Ridge Regression model as the most relevant model for the classification of user status for all tobacco products (results in Table 4). Since Kernel Ridge Regression projects variables into a higher Kernel space, it is not possible to interpret the weights of the variables in the model to perform feature selection. Thus, all variables are included in this model for our final prediction. For all model results, please check Table 5 in the appendix.

Table 4: Kernel Ridge Regression Results

User Status	Tobacco Product	MSE	AUC
Current User	Cigarettes	0.316	0.836
	E-cigarettes	0.202	0.883
	Traditional Cigar	0.092	0.883
	Cigarillo	0.112	0.904
	Filtered Cigar	0.060	0.885
	Hookah	0.079	0.875
	Pipe	0.030	0.822
	Snus	0.026	0.891
	Smokeless	0.082	0.875
Former User	Cigarettes	0.382	0.743
	E-cigarettes	0.357	0.785
	Traditional Cigar	0.294	0.842
	Cigarillo	0.295	0.856
	Filtered Cigar	0.256	0.832
	Hookah	0.323	0.797
	Pipe	0.226	0.871
	Snus	0.124	0.904
	Smokeless	0.205	0.841
Never User	Cigarettes	0.212	0.912
	E-cigarettes	0.297	0.868
	Traditional Cigar	0.279	0.868
	Cigarillo	0.255	0.900
	Filtered Cigar	0.266	0.850
	Hookah	0.311	0.833
	Pipe	0.231	0.872
	Snus	0.127	0.912
	Smokeless	0.215	0.873

7 Conclusion

Generally, machine learning algorithms such as Kernel Ridge, SVM perform better than traditional statistical models such as Logistic Regression and Lasso Linear Regression. More specifically, the Kernel Ridge Regression yields the best performance (highest AUC and similarly low test error compared with SVM and Random Forest) compared with other methods. In addition, the classification of modern and minor tobacco product use status based on the other products and sociodemographic factors yielded lower MSE than the classification of traditional tobacco products. In addition, the large class imbalances between current, former, and never users in modern and minor tobacco product use status classification may be causing a bias towards our low MSE results.

8 Limitations

Although our model results shows relatively low test error and high AUC, there are other factors we did not consider such as education level, Annual household income due to the great number of missing values of these variables in the dataset, and we could even tobacco related policy like Menthol Ban in several states. Besides that, our sample only included part of participants in the US, which might yield selection bias. However, we can multiply "weight" by using inverse probability weight method to each participant to represent the national level population. In addition, we did not have time to apply class balancing efforts to our dataset in the time allotted. Large class imbalances may have been a large contributing factor to the low MSE results achieved by the Kernel Ridge Regression method and balancing the classes may have resulted in a large improvement in true positive and true negative rate.

9 Contribution by Group Members

Each group member took charge of the classification of three different tobacco products. This included the model fitting, parameter tuning process and the evaluations for them. Authors 1 and 2 performed data preprocessing and summarized the results in the tables, while author 3 summarized the math formulas such as parameters and objective functions for the models. All group members worked on the report together. More specifically, Author 3 wrote Introduction, Motivation, Problem Statement, Related Work, and Descriptive Analysis and generated bar plots, author 1 wrote Data Preparation and Classification Methods and took charge of translating author 3's derived formulas into Latex and generating corresponding tables, and author 2 wrote Dataset, Evaluation, and Results. All authors wrote the abstract and the conclusion together.

10 References

[1] Neal L. Benowitz, M.D. Nicotine Addiction. New England Journal of Medicine.

[2] Graham W Warren, K Michael Cummings. Tobacco and lung cancer: risks, trends, and outcomes in patients with cancer. Soc Clin Oncol Educ Book.

[3] Rebecca Evans-Polce PhD, Phil Veliz PhD, etc. Trends in E-Cigarette, Cigarette, Cigar, and Smokeless Tobacco Use Among US Adolescent Cohorts, 2014–2018. American Journal of Public Health.

[4] Ali A, Hossain SM, Hovsepian K. mPuff: automated detection of cigarette smoking puffs from respiration measurements. 2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN), 2012.

[5] Cole CA, Thrasher JF, Strayer SM. Resolving ambiguities in accelerometer data due to location of sensor on wrist in application to detection of smoking gesture. IEEE, 2017: 489–92.

[6] Senyurek VY, Imtiaz MH, Belsare P, et al. A CNN-LSTM neural network for recognition of puffing in smoking episodes using wearable sensors. Biomed Eng Lett 2020;10:195–203.

[7] Visweswaran S, Colditz JB, O'Halloran P, et al. Machine learning classifiers for Twitter surveillance of Vaping: comparative machine learning study. J Med Internet Res 2020;22:e17478.

[8] Chu K-H, Colditz J, Malik M, et al. Identifying key target Audiences for public health campaigns: Leveraging machine learning in the case of Hookah tobacco smoking. J Med Internet Res 2019;21:e12443.

[9] Culotta A. Towards Identifying Leading Indicators of Smoking Cessation Attempts from Social Media. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016: 7–9.

[10] Kostygina G, Tran H, Shi Y, et al. 'Sweeter than a Swisher': amount and themes of little cigar and cigarillo content on Twitter. Tob Control 2016;25:i75–82.

[11] Patel J, Siddiqui Z, Krishnan A, et al. Leveraging electronic dental record data to classify patients based on their smoking intensity. Methods Inf Med 2018;57:253–60.

[12] Kim N, McCarthy DE, Loh W-Y, et al. Predictors of adherence to nicotine replacement therapy: machine learning evidence that perceived need predicts medication use. Drug Alcohol Depend 2019;205:107668.

[13] Mamoshina P, Kochetov K, Cortese F, et al. Blood biochemistry analysis to detect smoking status and quantify accelerated aging in smokers. Sci Rep 2019;9:142.

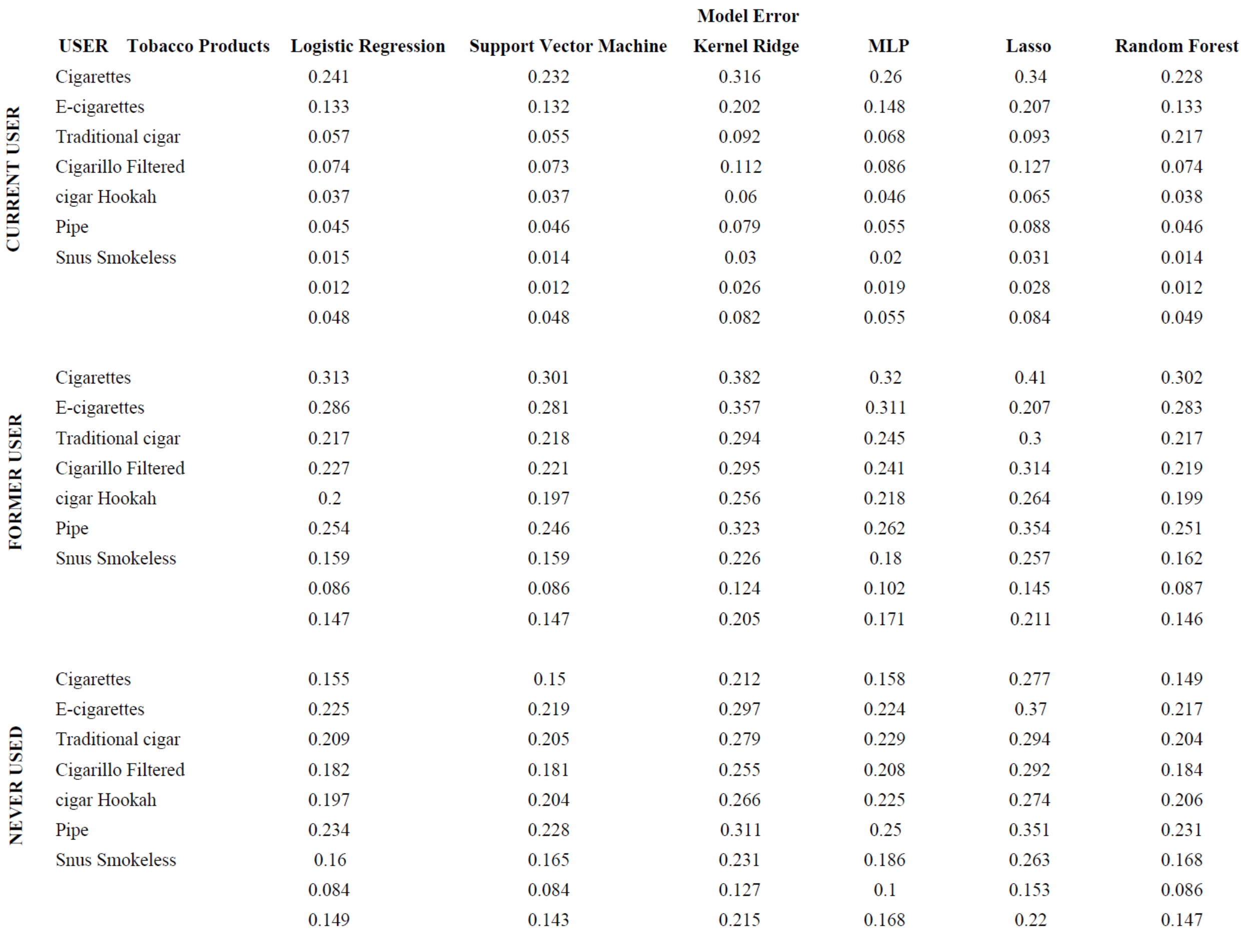
A Appendix

Table 5: Related Work

Topics	Publication Example	Objective	Methods
ML-powered technology to assist smoking cessation (n=22)	Ali A, Hossain SM etc. [4]	Automatically detecting smoking in the natural environment by developing mPuff, a model to automatically detect smoking puffs from respiration measurement.	Semi-supervised support vector machine (SVC) that improves the accuracy of detecting smoking puffs by making use of the self-report smoking.
	Cole CA, Thrasher JF etc. [5]	Presented an investigation into the ambiguities in accelerometer data that arise due to the position of the smart watch on a person's wrist to detect smoking gestures.	Artificial Neural Networks (ANNs)
	Senyurek VY, Imtiaz MH. [6]	Proposed a novel algorithm for automatic detection of puffs in smoking episodes by using a combination of Respiratory Inductance Plethysmography and Inertial Measurement Unit sensors.	Convolutional neural networks (CNN) + Long Short-Term Memory (LSTM) network layers (Deep learning)
Content analysis of tobacco on social media (n=32)	Visweswaran S, Colditz JB etc. [7]	This study aims to derive and evaluate traditional and deep learning classifiers that can identify tweets relevant to vaping, tweets of a commercial nature, and tweets with provape sentiments.	1. A set of 4000 tweets was selected, and each tweet was manually annotated for relevance (vape relevant or not), commercial nature (commercial or not), and sentiment (pro-vape or not). 2. Using the annotated data (labeled data), the author derived traditional classifiers (random forest, multinomial naive Bayes, etc.) 3. Using the annotated data and unannotated data to derive deep learning classifiers (LSTM, CNN etc.) 4. Compare the accuracy of both classifiers.
	Chu K-H, Colditz J. [8]	1. Confirm previous research showing positively skewed Hookah tobacco smoking (HTS) sentiment on Twitter. 2. Identify individuals who exhibit mixed opinions about HTS via the Twitter platform and therefore represent key audiences for intervention.	1. Labeled sentiment towards HTS for the training dataset by using Natural language processing software to extract linguistic features. 2. Test model in the testing dataset. 3. Identify people with both positive and negative attitude towards HTS to do intervention.
	Culotta A. [9]	Identify leading indicators of smoking cessation attempts from Twitter data. Specifically, identify linguistic patterns that are characteristic of smokers who are likely to attempt to quit smoking in the near future.	1. Manually labeled tweets related to smoking attempts. 2. Trained model using supervised learning (Logistic Regression). 3. Tested data.
	Kostygina G, Tran H. [10]	Identify the amount or content of little cigar and cigarillo (LCC) messages users see or share on social media.	1. Collect LCC related data from twitter. 2. Tweets were coded for promotional content, brand references, co-use with marijuana and subculture references. 3. Keywords algorithms were used to classify commercial or non-commercial.

Smoker status classification from narrative clinical texts (n=6)	Patel J, Siddiqui Z etc. [11]	To determine patients' detailed smoking status based on smoking intensity from the EDR.	SVM, Random Forest
Tobacco-related outcome prediction using administrative, survey or clinical trial data (n=14)	Kim N, McCarthy DE. etc. [12]	Identify pre-quit predictors of nonadherence to smoking cessation medication, which may help explain nonadherence and suggest tailored interventions to address it.	Secondary classification decision tree analyses in a 2-arm RCT of Recommended Usual Care (R-UC) VS. Abstinence-Optimized Treatment (A-OT)
	Mamoshina P, Kochetov K etc. [13]	1. Use blood biochemistry and cell count results to predict smoking status and then use the smoking status to assess the effect on biological aging. If this prediction has high accuracy, self-reported smoking could be replaced by blood biochemistry and cell count results.	Supervised deep learning.

Table 6: MSE Results



				Model Error			
USER	Tobacco Products	Logistic Regression	Support Vector Machine	Kernel Ridge	MLP	Lasso	Random Forest
CURRENT USER	Cigarettes	0.241	0.232	0.316	0.26	0.34	0.228
	E-cigarettes	0.133	0.132	0.202	0.148	0.207	0.133
	Traditional cigar	0.057	0.055	0.092	0.068	0.093	0.217
	Cigarillo Filtered	0.074	0.073	0.112	0.086	0.127	0.074
	cigar Hookah	0.037	0.037	0.06	0.046	0.065	0.038
	Pipe	0.045	0.046	0.079	0.055	0.088	0.046
	Snus Smokeless	0.015	0.014	0.03	0.02	0.031	0.014
		0.012	0.012	0.026	0.019	0.028	0.012
		0.048	0.048	0.082	0.055	0.084	0.049
FORMER USER	Cigarettes	0.313	0.301	0.382	0.32	0.41	0.302
	E-cigarettes	0.286	0.281	0.357	0.311	0.207	0.283
	Traditional cigar	0.217	0.218	0.294	0.245	0.3	0.217
	Cigarillo Filtered	0.227	0.221	0.295	0.241	0.314	0.219
	cigar Hookah	0.2	0.197	0.256	0.218	0.264	0.199
	Pipe	0.254	0.246	0.323	0.262	0.354	0.251
	Snus Smokeless	0.159	0.159	0.226	0.18	0.257	0.162
		0.086	0.086	0.124	0.102	0.145	0.087
		0.147	0.147	0.205	0.171	0.211	0.146
NEVER USED	Cigarettes	0.155	0.15	0.212	0.158	0.277	0.149
	E-cigarettes	0.225	0.219	0.297	0.224	0.37	0.217
	Traditional cigar	0.209	0.205	0.279	0.229	0.294	0.204
	Cigarillo Filtered	0.182	0.181	0.255	0.208	0.292	0.184
	cigar Hookah	0.197	0.204	0.266	0.225	0.274	0.206
	Pipe	0.234	0.228	0.311	0.25	0.351	0.231
	Snus Smokeless	0.16	0.165	0.231	0.186	0.263	0.168
		0.084	0.084	0.127	0.1	0.153	0.086
		0.149	0.143	0.215	0.168	0.22	0.147

Table 7: AUC Results

				Model AUC			
	Tobacco Products	**Logistic Regression**	**Support Vector Machine**	**Kernel Ridge**	**MLP**	**Lasso**	**Random Forest**
CURRENT USER	Cigarettes	0.735	0.752	0.836	0.721	0.806	0.757
	E-cigarettes	0.596	0.6	0.883	0.637	0.858	0.534
	Traditional cigar	0.596	0.6	0.883	0.537	0.5	0.595
	Cigarillo	0.659	0.672	0.904	0.654	0.894	0.595
	Filtered cigar	0.544	0.513	0.885	0.583	0.863	0.529
	Hookah	0.534	0.519	0.875	0.57	0.832	0.515
	Pipe	0.507	0.541	0.822	0.538	0.5	0.534
	Snus	0.544	0.5	0.891	0.576	0.847	0.509
	Smokeless	0.562	0.566	0.875	0.581	0.796	0.562
FORMER USER	Cigarettes	0.619	0.619	0.743	0.615	0.685	0.62
	E-cigarettes	0.65	0.68	0.785	0.647	0.762	0.691
	Traditional cigar	0.719	0.719	0.842	0.704	0.829	0.722
	Cigarillo	0.741	0.753	0.856	0.729	0.841	0.722
	Filtered cigar	0.624	0.612	0.832	0.628	0.828	0.617
	Hookah	0.665	0.669	0.797	0.66	0.775	0.668
	Pipe	0.699	0.681	0.871	0.688	0.858	0.691
	Snus	0.64	0.624	0.904	0.618	0.896	0.668
	Smokeless	0.61	0.617	0.841	0.628	0.832	0.614
NEVER USED	Cigarettes	0.769	0.773	0.912	0.764	0.891	0.76
	E-cigarettes	0.775	0.782	0.868	0.768	0.854	0.703
	Traditional cigar	0.77	0.777	0.868	0.75	0.855	0.778
	Cigarillo	0.815	0.817	0.9	0.789	0.891	0.778
	Filtered cigar	0.674	0.672	0.85	0.674	0.842	0.675
	Hookah	0.729	0.737	0.833	0.717	0.81	0.73
	Pipe	0.725	0.705	0.872	0.7	0.859	0.703
	Snus	0.709	0.71	0.912	0.7	0.905	0.723
	Smokeless	0.701	0.718	0.873	0.704	0.864	0.698