Time-Course Cardiovascular Risk Prediction Using Machine Learning Models

Tien D. Pham **YuYing Chen** **Qingyang Feng** **Xianyuan Zeng**

Abstract

Using the data obtained from the Framingham Heart Study, this project compares the usage of the time-dependent Cox PH survival model against time-independent machine learning (ML) algorithms: Naïve Bayes, Random Forest (RF) and Logistic Regression (LR) models, for predicting clinical cardiovascular conditions on longitudinal datasets. We found that despite being from the same class of probabilistic models, the hazard ratios estimated by the Cox model are related, but not comparable to the likelihoods from the ML models. Moreover, while noticeably impacted by the characteristics of longitudinal studies such as data censoring, dependent observations and severe class imbalance, both RF and LR models were able to identify a similar set of risk factors for each cardiovascular condition compared to the Cox model. This suggests that ML algorithms can reinforce and complement the diagnostic and prognostic capacity of statistical survival analysis.

1 Introduction

Coronary heart disease (CHD) has long been a key area of focus in the discussion of public health, particularly in the predictive healthcare field. As such, numerous studies have been conducted to identify risk factors leading to the onset of different cardiovascular conditions, ranging from using classic statistical procedures such as regression models to modern machine learning (ML) algorithms [1][2][3]. Although these techniques aim to provide risk predictions using data from a specific timepoint, the time-course effects of the predictors (i.e., how these predictors interact with each other over time) on the target events should not be overlooked from an etiological perspective.

The need to include time-dependent risk factors can be aptly addressed using survival analysis techniques, most notably the time-dependent Cox proportional hazards (PH) regression model [4]. As a robust semi-parametric model, Cox PH has been a staple method in survival analysis where the focus is emphasised on the expected duration of time until an event happens (*regression* task, mainly used in prognosis) rather than whether such an event will happen (*classification* task, mainly used in diagnosis). By taking the proportion of the estimated incident counts for the interested interval, the Cox PH can be also be used as a diagnostic classifier. It has been shown to marginally outperform Logistic Regression and other regression techniques when the time-to-event information is available and there is time-related missing data (i.e., **censoring**) [5].

While there have been a number of studies that compare the Cox PH model against ML models in predicting cardiovascular conditions, very few of them took into account the difference in the diagnostic and prognostic use cases of these two techniques for different types of diseases. This project seeks to fill in this gap by comparing how the two techniques can be applied to predict whether a patient has developed a specific type of CHD given the observed risk factors, and investigating the condition-specific risk factors identified by each model.

Preprint. Under review.

Matrix X | Target y | Not used in ML

ID (4434)	PREDICTORS (19)	TIME	PERIOD	EVENT (8)	TIME TO FIRST EVENT
	▪ Demographic RF ▪ Behaviour RF ▪ Medical History ▪ Baseline measure	Days since start of the follow-up	Follow up number: 1, 2, or 3	1 if an event occurs between the end of the previous period & the current period **or if the event occurs before current period**, 0 otherwise or after the study finishes	Days since start until the **first** event occurrence

Figure 1: Summary of Framingham dataset structure.

2 Dataset

The examined dataset is a subset of the data collected as part of the Framingham Heart Study (FHS) and includes adjudicated event data on 4,434 participants, which is obtained for educational purposes with permissions from the National Heart, Lung, and Blood Institute (NHLBI) of the United States of America. The FHS was a longitudinal prospective study of etiology of cardiovascular disease among population living in the community of Framingham, Massachusetts. The participants were examined biennially and continuously followed through regular surveillance for occurrence of Angina Pectoris, Myocardial Infarction, Stroke, Hypertension and Death from any causes [6]. The current dataset is provided in a .csv format in a longitudinal form, with a total of 11,627 observations on the 4434 participants, of which 56.2% were women (n=2490) and 43.8% were men (n=1944). Each observation has 21 risk factors (predictors) and 8 target events (targets) including 3 umbrella events: MI_FCHD, ANYCHD, and CVD (Figure 1). The mean age was 54.8 years (standard deviation (SD)=9.6). 52.6% of the patients (n=2334) were smokers, 59.3% had history of hypertension (n=2631) and 7.6% had diabetes (n=338). The mean total cholesterol was 241.3 (SD=45.4), the mean systolic blood pressure was 136.4 mmHg (SD=22.8), and the mean diastolic blood pressure was 83.1 mmHg (SD=11.7).

The dataset also exhibits **censoring** - a key characteristic of longitudinal data in which the observations are not uniformly observed for all subjects due to some subjects leaving the study prior to experiencing an event (either due to dropout or death). In this project, we investigated two common approaches to censored data: through survival analysis which is robust towards censoring, and through treating each observation as independent of the patient before training ML models, effectively removing the time-course nature of the longitudinal dataset [5] and the need to use the *Time to first event* attributes.

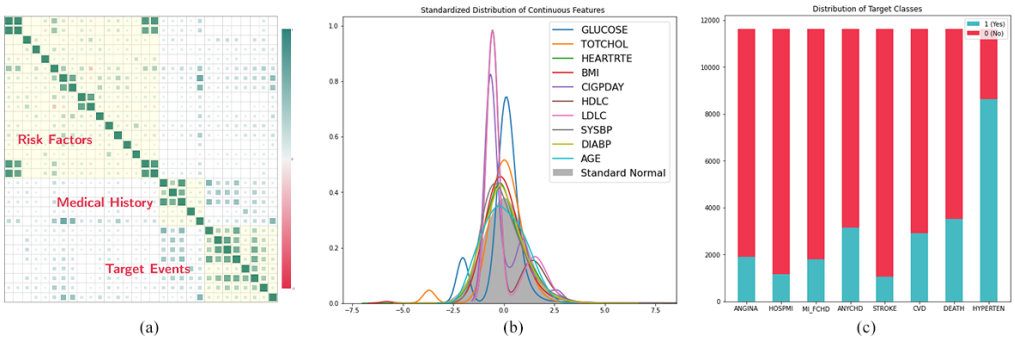
3 Methods

In this section, we describe how we prepared the FHS dataset and the hyperparameters with which we trained our ML models and Cox PH models. We also provide overviews of the different metrics and techniques used for model evaluation and inference.

3.1 Data preparation

In addition to the typical characteristics of longitudinal data, the current FHS dataset is also subject to the curse of dimensionality, especially multicollinearity, a major challenge in working with medical datasets as discussed in [7]. With 19 original predictors and 8 target events, it warrants meticulous and uniform data preparation across different models and target events so that the model results are comparable and interpretable.

Data pre-processing Overall, the dataset well-formed and thoroughly documented by NHLBI [8], with only missing-at-random data due to the censoring issue identified above. Specifically, measurements for total cholesterol, BMI, glucose as well as information about education level, number of cigarette per day and use of anti-hypertensive medications were randomly missing. For these essential and irremovable variables, we redressed this issue through simple zero imputations (labelled as *Original*) and column mean imputations (labelled as *Filled*). Without further medical domain expertise, it was challenging to provide better estimates considering the high complexity cost of more advanced imputation techniques. For the Density Lipoprotein Cholesterol variables (*HDLC*



(a) (b) (c)

Figure 2: (a) Correlation matrix with highlighted variable groups; (b) Standardised Continuous Feature Distribution; (c) Target Class Distribution.

and *LDLC*) which are conditionally available for patients who had a 3rd follow-up according to the data dictionary, zero imputations were used throughout for both *Original* and *Filled* variants of the dataset, as mean imputation would not honour the conditional and would change the meaning of these variables.

Note that missing data imputations were only applied to the predictor features in the dataset (Figure 1) for the ML model pipeline, where each observation is treated as independent of time and individual patient. This would enable models such as Naive Bayes and Logistic Regression from the linear classifier family to be fitted. In comparison, the Cox PH algorithm can robustly handle censored data and therefore does not require imputations [9].

Feature selection A correlation matrix was generated from the *dython* package to address the issue of multicollinearity (Figure 2a). Different correlation metrics were calculated for different combination of data types: Pearson's r for continuous variables, Crammer's V for nominal variables, and correlation ratio for continuous variable versus nominal variable [10]. We did not remove multicollinearity in the risk factors that were measured in pairs, such as *HDLC* and *LDLC* or *SYSBP* and *DIABP*, as they are medically complementary of each other despite the strong correlation. However, there are also overlapping in the definitions of prevalent medical history. We thus excluded prevalent Angina Pectoris and Myocardial Infarction from the feature set as they have been accounted for under the prevalent Coronary Heart Disease indicator.

A density plot for all continuous variables was also used as a sanity check for normality assumption (Figure 2b). The multimodal variables coincide with the variables that were imputed, which was an expected trade-off from using single-value imputations. While no variables were further dropped, this normality check suggested that the preprocessed dataset may not perform well if a model requires normally distributed feature (e.g., Gaussian Naive Bayes).

Imbalanced target distribution Figure 2c summarises the class distribution for each target event. With the exception of Hypertension (which is forewarned to be highly subject to misclassification in the FHS documentation), all cardiovascular conditions have a low prevalence, with an average of 19.03%. Since it is well-known that this imbalance in class distribution complicates the training and evaluation of most ML classifiers [11], we addressed it at two touchpoints in our experimental pipeline. Firstly, we created two additional variants of the dataset, *SMOTE* and *Filled SMOTE* based on the Synthetic Minority Over-sampling Technique [12] to balance the target distribution for each event. Secondly, the evaluation and tuning metrics were chosen such that they are not vastly affected by class imbalance (see **Evaluation metric** in 3.2).

3.2 Machine Learning model descriptions

For each of the four dataset variants (original, filled, SMOTE, filled SMOTE), we split it into training and testing sets with 70% and 30% of the data respectively, and trained the following models to predict each one of the 8 target events using their default settings in the *scikit-learn* library. In total, 96 machine learning models were fitted (Figure 3), and the results were averaged over all targets

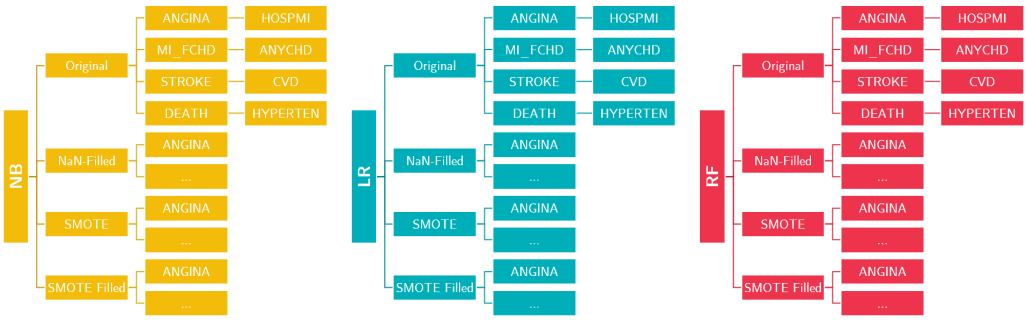


Figure 3: Design of Experiment (3 models x 4 train sets x 8 target events).

using the test sets. We then proceeded to tuning the best classifiers on the *class_weight* parameter to combat the abovementioned imbalanced class problem.

Mixed Naive Bayes (NB) NB is a probabilistic classifier which estimates the posterior probability of occurrence as the product of the prior class distribution, the distribution of discrete features and the Gaussian-fitted distribution of continuous features. Although the selected feature set satisfies the feature independence assumption, most continuous features were not sufficiently normal for the likelihoods to be accurately estimated (Figure 2b). Furthermore, the imbalanced prior distribution may also affect posterior estimation. For these drawbacks, we only consider NB as a baseline model for ML performance. We combined both *CategoricalNB()* and *GaussianNB()* from *scikit-learn* as our own NB implementation.

$$p(C|x_1, ..., x_{18}) \propto p(C) \prod_{i=1}^{18} x_i p(x_i|C)$$

Logistic Regression (LR) Similar to NB, LR is a probabilistic classifier which estimates the probability of occurrence by linearly assigning a weight w_i to each predictor x_i. One distinct advantage of LR over NB is that it does not require x_i to be normal, which results in more accurate predictions when all 18 features are included and treated as numerical variables. While we have addressed the multicollinearity issue in pre-processing, the longitudinal nature of the dataset may violate the LR assumption of independent observations. Fortunately, Allison (2010) has noted that there is no inflation of test statistics resulting from a lack of independence for a dataset with multiple records for intervals within each individual [13]. As such, the time-independent approach (treating all records as conditionally independent) is sufficiently suitable to fit a LR. Our *scikit-learn* implementation used the default setting for binary classification, and all models fully converged when the *max_iter* parameter was set to 10000. We also noted that LR is sensitive to class imbalance, and thus followed up training with grid-search tuning a weighted LR model.

$$p(C|x_1, ..., x_{18}) = \frac{1}{1 + \exp \sum_{i=1}^{18} -w_i x_i}$$

Random Forest (RF) RF is an ensemble-based technique using multiple decision trees to improve performance and lower biases. Although the decision tree algorithm only produces a single class prediction and thus is not probabilistic like NB and LR, the *scikit-learn* implementation of RF (with 100 *n_estimators* using the default *gini* split criterion) combines single trees by averaging the fraction of training samples of the same class in a leaf as a *probability*, instead of letting each classifier vote for a single class. Like LR, RF is subject to class imbalance, which we also addressed by tuning the *class_weight* parameter.

Evaluation metric While the F1-score is widely used as a default metric for imbalanced class problem, there has been a strong advocacy for using the Matthews correlation coefficient (MCC) as a more holistic and reliable metric for binary classifications [14]. Taking values between -1 (completely wrong predictions) and 1 (completely correct predictions), the MCC summarises the information

about the observed and predicted classes in a confusion matrix without emphasising either false negatives or false positives as this may lead to overoptimistic inflated results.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

In addition to MCC as a novel evaluation metric, we also adopted the more familiar Average Precision (AP) score, which summarises a precision-recall curve (PRC) as the weighted mean of precisions achieved at each threshold [15]. The PRC is an alternative to the receiving operating characteristic curve (ROC), and has been shown to be more robust to imbalanced class problems. Other than visualising the PRC, the AP scores were also used to evaluate hyperparameter tuning and feature importance, thanks to its integrated support in the *scikit-learn* library.

3.3 Cox Proportional Hazard Model (Cox PH)

As briefly introduced, the Cox PH model has been successfully used to predict event occurrence by modelling the time-to-event variable from the predictors (also known as **covariates** in the context of survival analysis). It assumes that the covariates have a multiplicative effect on the **hazard function** (i.e., the risk of having the target event). Under the time-dependent Cox PH model, the hazard function for subject i given their observations $\mathbf{X}_i$ at time t is given by:

$$\lambda(t|\mathbf{X}_i(t)) = \lambda_0(t) \exp(\beta \mathbf{X}_i)$$

Note that here $\mathbf{X}_i$ excludes the *Time* feature; and as such, the coefficient matrix β only have 17 terms, instead of 18 terms like in the Naive Bayes and Logistic Regression models. The likelihood of the event to occur for subject i at time t_i can then be written as a proportion involving hazard functions:

$$p_i(C|t_i, \mathbf{X}) = \frac{\lambda(t_i|\mathbf{X}_i)}{\sum_{j:t_j \geq t_i} \lambda(t_i|\mathbf{X}_j)} = \frac{\lambda_0(t_i) \exp(\beta \mathbf{X}_i)}{\sum_{j:t_j \geq t_i} \lambda_0(t_i) \exp(\beta \mathbf{X}_j)} = \frac{\exp(\beta \mathbf{X}_i)}{\sum_{j:t_j \geq t_i} \exp(\beta \mathbf{X}_j)}$$

Unlike the predictions made by ML models, this likelihood is computed using not only the observations from subject i, but also from other subjects j for whom the event has not occurred before time t_i. Since the predictions are relative to the population covariates that depend on the predicting instance, ML performance metrics such as MCC and AP cannot be used to evaluate Cox PH models. Instead, for each target event, we fitted the model on the whole dataset and evaluated the goodness-of-fit using the Likelihood ratio test and the Wald test using the *survival* library from [16].

3.4 Model inference

Although the probabilities calculated for each model are semantically different, the set of important features used to estimate such probabilities remains comparable across different models. For LR and RF, feature importance is measured by the loss in AP score if the observation vector for that feature is stochastically permuted [17]. The larger the change in the performance, the more important the feature is. For Cox PH models, feature importance is directly interpreted from the fitted coefficient vector β. Our analysis focused on both the most important features used to make predictions from the three models for each target event, and the most important risk factors that contribute to the prognosis of cardiovascular diseases.

4 Results

4.1 ML models performance

Training sets Figure 4 implies that as a baseline model, the results for NB are consistent regardless of train set variants. In contrast, both LR and RF significantly worsen in performance after SMOTE algorithm was performed to balance the target classes. There is no difference between zero imputation and mean imputation variants for both the original and SMOTE sampled datasets. As both LR and RF recorded the same average MCC for the *Original* and *Filled* training sets, we decided to use the latter variants for hyperparameter tuning.

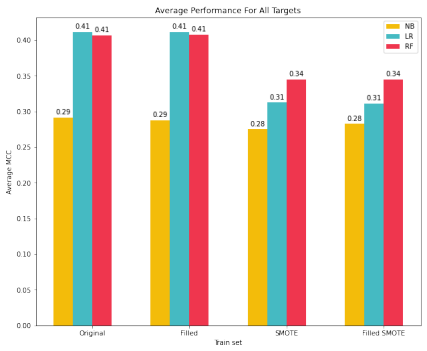


Figure 4: Average MCC across all events for different train sets of NB (yellow), LR (teal), and RF (red).

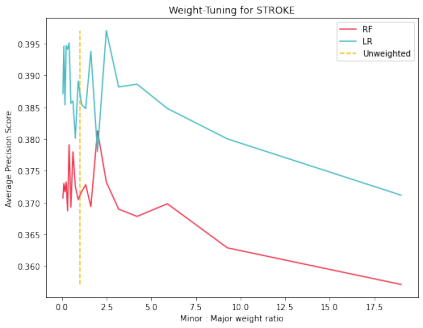


Figure 5: Class weight tuning for LR (teal) and RF (red) on STROKE. The yellow baseline denotes the unweighted (1:1) ratio.

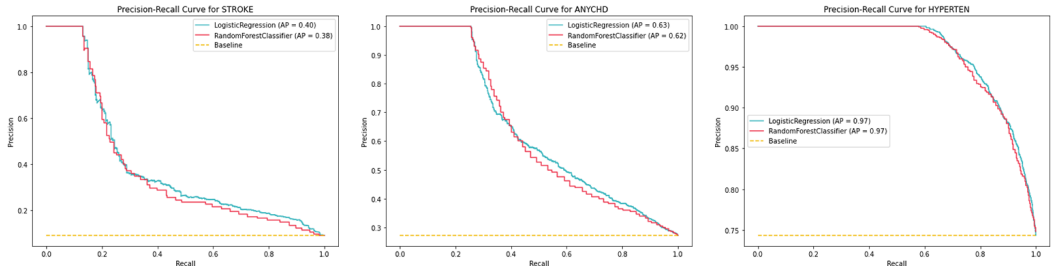


Figure 6: Precision-Recall curves comparison between LR (teal) and RF (red) against baseline (yellow) for (a) STROKE, (b) ANYCHD, (c) HYPERTEN.

Class weight tuning Figure 5 illustrates the results from grid search with 3-fold cross validation in mean AP score for different class weight ratios for both LR and RF on *STROKE*, the target event with has the lowest prevalence among all events. The best minor-to-major ratio for LR was approximately 2.6, compared to 2.0 for RF. However, since the improvement was only 0.005 and the weight ratio was only applicable to the *STROKE* event, we decided to keep all models as unweighted for further analysis.

Target events There was no discernible difference between the two unweighted models when tested on a similar test set of 3488 instances as seen in Figure 6, irrespective of the target event, both in terms of the PRC shape and the AP score. In contrast, there was a clear improvement in performance between different events as the distribution of class becomes more balanced, suggesting that the implemented techniques to overcome class imbalance were unsuccessful.

4.2 Cox PH regression result

At 13 degrees of freedom, both the Wald test and the likelihood ratio test for all the Cox PH models fitted on different target events were highly significant ($p < .001$). Figure 7 gives an example of how the hazard functions can be used to predict the prognosis of different cardiovascular conditions for two female, smoking and non-diabetic patients. Note that while the Cox PH model considers each patient's records as a series of measurements, its prediction is based on proportional hazards. Thus it will return 3 hazard functions for a patient with 3 follow-ups, each of which corresponds to the prognosis of that patient at that specific follow-up. In our example, patient 10552 was older and had a history of hypertension, resulting in her cumulative risk of developing all three conditions being significantly higher, with a near 30% chance of death between Period 2 and Period 3. This was the actual outcome recorded in the FHS. For patient 11252, her Period 2 prognosis was the best out of three follow-up periods, which was due to a record of lower blood pressure and heart rate. These two

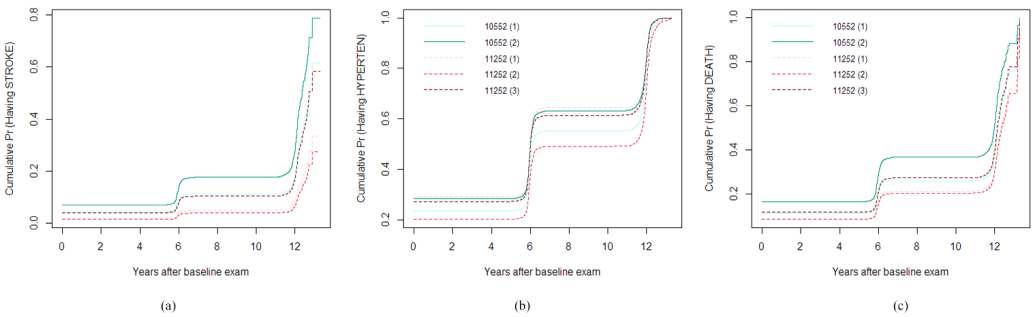


Figure 7: Cox Cumulative event probability prediction for patient 10552 (teal) and 11252 (red) on 3 follow-ups (from lighter to darker) for (a) STROKE, (b) HYPERTEN, (c) DEATH.

Table 1: Feature Importance for each model using coefficients for Cox (CPH) and AP drop-out loss for LR and RF. Features with highest importance in each column are bolded. Features with zero importance are omitted.

	ANGINA			ANYCHD			CVD			DEATH			HOSPMI			HYPERTEN			MI_FCHD			STROKE		
	CPH	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**	**CPH**	**LR**	**RF**
AGE	-.039	.255	.003	-.028	.236	.005	-.012	.248	.026	.011	**.358**	**.152**	-.038	.250	.002	-.046	.087	.002	-.025	.259	.004	.005	**.299**	**.040**
BMI	.032	.260	.006	.026	.231	.008	.023	.218	.005		.181	.002	.010	.255	.003	.011	.087	.006	.016	.250	.008	.020	.205	
BPMEDS	-.042	.260		-.103	.239		-.155	.220		-.149	.181		-.188	.257		-.302	.087		-.113	.247		-.093	.200	
CIGPDAY	-.003	.256		.003	.232		.006	.218	.001	.009	.182	.003	.007	.255	.001	-.004	.086		.007	.247	.002	.009	.204	
CURSMOKE		.255			.231			.221			.180			.262			.086			.259			.205	
DIABETES	-.015	.262		.159	.239		.313	.225		.255	.190		.353	.260		-.199	.086		.356	.264		.370	.209	
DIABP		.251	.003		.234	.003		.216	.006		.181	.003		.249	.002		.088	.013		.248	.002		.205	.002
EDUC	-.064	.255	.002	-.117	.231		-.138	.221		-.172	.180	.001	-.084	.256		-.115	.088		-.086	.246		-.151	.207	
GLUCOSE		.252	.001		.231	.001		.218	.001		.180	.003		.256	.001		.086	.001		.247	.002		.205	
HDLC		.274			.239			.232			.181			.263			.089			.259			.204	
HEARTRTE	-.009	.257	.001	-.007	.229		-.006	.217	.002	-.001	.179	.001	-.008	.254		-.002	.086		-.004	.247	.001	-.007	.205	
LDLC		.265			.234			.221			.180			.276			.086			.253			.205	
PREVCHD	**1.880**	**.422**	**.061**	**1.285**	**.339**	**.067**	.719	.259	.020	.336	.187	.002	**1.526**	**.341**	.030	-.013	.086		**1.185**	.313	.026	.141	.203	
PREVHYP	-.190	.273		-.166	.243		-.019	.238	.001	-.143	.183	.001	-.248	.275		**.233**	**.297**	**.178**	-.169	.253		.185	.229	
PREVSTRK	-.286	.255		-.044	.231		**.890**	.245	.010	**.359**	.187		-.202	.255		.043	.086		.083	.248		**1.935**	.249	.022
SEX	-.222	.265	.002	-.485	.277	.035	-.626	**.266**	**.041**	-.490	.209	.022	-1.050	.304	**.041**	.002	.089		-.948	**.314**	**.059**	-.135	.205	
SYSBP	.009	.260	.005	.012	.246	.021	.012	.226	.018	.013	.197	.017	.014	.307	.013	.012	.161	.067	.015	.265	.015	.014	.221	.007
TIME		.271	.005		.265	.008		.257	.013		.256	.041		.255	.001		.102	.007		.254	.001		.261	.002
TOTCHOL	.004	.271	.013	.004	.236	.011	.004	.220	.006	.002	.180	.002	.006	.264	.003	.002	.086	.001	.006	.255	.009		.206	

factors are reported to relate to better prognosis for these events according to the model coefficients (see Table 1).

4.3 Model inference

As in Table 1, all 3 models placed strong emphases on similar sets of high-risk factors for each target event, namely prevalent CHD for *ANGINA, ANYCHD, HOSPMI*, sex for *CVD, MI_FCHD*, age and prevalent stroke for *DEATH, STROKE*. There were more overlapping in feature importance for LR and RF, which may explain the similarity in performance across different targets. Note that while the loss in AP score is an effect size measure, the Cox coefficients are signed indicators for either positive protective factors or negative risk factors. Thus the absolute values of these feature importance measures cannot be compared cross-models.

5 Discussion

5.1 Model performance

We obtained similar results with previous works with the comparable performance between LR and RF albeit using different metrics. From a predictive modelling standpoint, the areas under the PRC, which are estimated by AP scores, were not high (ranging from 0.35 to 0.66 except for *HYPERTEN*). This was primarily due to the minimal efforts to engineer features with higher predictive capacity and lack of more extensive hyperparameter tuning. However, considering the coverage of the design of experiment which spans all 8 CHD conditions instead of just 1 target event like most previous results focused on, and that we only used a minimally pre-processed dataset which does not deviate significantly from the original FHS dataset, our results can be used as a benchmark to compare with future published results on more refined models using the same sets of metrics that circumvent class imbalance (i.e., MCCs and AP scores).

Neither the use of SMOTE resampling nor increasing the weights of the minority class successfully reduced the effect of class imbalance for both models. Specifically for the prior, we suspected that the traditional SMOTE algorithm could not simulate the longitudinal relationship between the instances, and thus resulted in structural changes in the dataset that deteriorated performance. This is in congruence with the results from [18], which has found that SMOTE does not attenuate the bias towards the classification when data are high-dimensional like the current dataset. In such cases, random undersampling could be explored as an alternative solution.

5.2 Prognostic versus Diagnostic use case

The design of our experiments has also drawn a distinct line between the use case and trade-offs between ML models and Cox PH models for different predictive healthcare purposes. ML model outputs are intuitively clearer to interpret than Cox PH outputs, since they are diagnostic predictions using a single patient's observation at a single time point. In contrast, Cox PH outputs compare the patient's observation with other patients, and do not provide a probabilistic prediction. The proportional hazard ratios instead estimate whether the current patient is more or less susceptible to the condition than others.

On the other hand, ML algorithms cannot model the aforementioned probability with respect to time (i.e., the time-course prognosis of the condition). While we could theoretically transformed the dataset to a long format wherein each patient instance has the records of all the follow-ups to fit the models, the censoring problem would still persist. For this use case, the time-dependent Cox PH, as a semi-parametric statistical model, is more suitable where the proportions provide an intuition solution to the challenge of having multiple records per patient, as explained with the example in Section 4.2.

5.3 Risk factors

Despite the difference in use case suitability, both ML models and Cox PH models were found to identify a similar set of risk factors associated with each type of cardiovascular condition. Aside the obvious important features such as medical history and age, the models were able to discern some interesting condition-specific risk factors, such as men's higher risk of Myocardial Infarction than women. However, all three models seemed to overemphasise the predictive power of medical history (i.e., prevalence of CHD), and as such did not highlight the roles of behavioural and demographic factors such as smoking and education level in the development of CHD. Specifically for the RF models which use fewer features to predict than LR (which did not use Lasso regression) as part of the algorithm, the prevalent variables can be up to 20 times more important than other variables.

Another issue with using highly correlated prevalent variables is the overinflation of Cox PH coefficients [19]. This results in contradictory regression results, where having history of one CHD decreases the hazard function for another condition, as in the case of Angina Pectoris. While this issue does not necessarily mean that the model is incorrect, it reduces the interpretability of the Cox coefficients, which is arguably one of the advantages of the Cox PH models over ML models.

5.4 Future improvements

Our current analysis has provided some benchmark results from evaluating ML models and Cox PH models for specific CHDs using appropriate metrics for imbalanced class, without rigorous model tuning. As such, it warrants the exploration of more fine-tuned and state-of-the-art models, such as neural networks or time-varying covariates Cox PH models. Alternatively, it would be interesting to consider a different set of predictors without previous medical conditions to see if the performance still holds for models with only demographic and behavioural risk factors. Finally, the major issue of class imbalance has yet to be addressed, for which other sampling techniques that can overcome the time-course challenge of the dataset should be considered.

References

[1] Juan-Jose Beunza et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)". In: *Journal of biomedical informatics* 97 (2019), p. 103257.

[2] Emelia J Benjamin et al. “Independent risk factors for atrial fibrillation in a population-based cohort: the Framingham Heart Study”. In: *Jama* 271.11 (1994), pp. 840–844.

[3] Tamara Harris et al. “Body mass index and mortality among nonsmoking older persons: the Framingham Heart Study”. In: *Jama* 259.10 (1988), pp. 1520–1524.

[4] Lloyd D Fisher and Danyu Y Lin. “Time-dependent covariates in the Cox proportional-hazards regression model”. In: *Annual review of public health* 20.1 (1999), pp. 145–157.

[5] Julius S Ngwa et al. “A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study”. In: *BMC medical research methodology* 16.1 (2016), p. 148.

[6] *About FHS | Framingham Heart Study*. URL: `https://framinghamheartstudy.org/fhs-about/`.

[7] Choong Ho Lee and Hyung-Jin Yoon. “Medical big data: promise and challenges”. In: *Kidney research and clinical practice* 36.1 (2017), p. 3.

[8] *Teaching Datasets - Public Use Datasets, National Heart, Lung, & Blood Institute*. URL: `https://biolincc.nhlbi.nih.gov/teaching/`.

[9] Bradley Efron. “The efficiency of Cox’s likelihood function for censored data”. In: *Journal of the American statistical Association* 72.359 (1977), pp. 557–565.

[10] *“dython”: A set of data tools in Python*. URL: `https://pypi.org/project/dython/`.

[11] Nathalie Japkowicz and Shaju Stephen. “The class imbalance problem: A systematic study”. In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[12] Nitesh V Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[13] Paul D Allison. *Survival analysis using SAS: a practical guide*. Sas Institute, 2010.

[14] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21.1 (2020), p. 6.

[15] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[16] Terry M Therneau. *A Package for Survival Analysis in R*. R package version 3.2-7. 2020. URL: `https://CRAN.R-project.org/package=survival`.

[17] Przemyslaw Biecek. “DALEX: Explainers for Complex Predictive Models in R”. In: *Journal of Machine Learning Research* 19.84 (2018), pp. 1–5. URL: `https://jmlr.org/papers/v19/18-416.html`.

[18] Lara Lusa et al. “Improved shrunken centroid classifiers for high-dimensional class-imbalanced data”. In: *BMC bioinformatics* 14.1 (2013), p. 64.

[19] Kristina P Vatcheva et al. “Multicollinearity in regression analyses conducted in epidemiologic studies”. In: *Epidemiology (Sunnyvale, Calif.)* 6.2 (2016).