Final Report for CIND820:

Predicting Survivability of Colorectal Cancer Using a Cluster-Then-Predict Method

Ryerson University

Contents

Introduction

Cancer survivability has been predicted using various classification machine learning models. In the medical community, a patient is considered in remission or cured if signs of cancer are diminished after 5 years, therefore, cancer survivability is usually posed as a binary classification problem. In models where regression is used to predict a specific survival time, it is usually used after data is separated into patient cohorts by known survival periods. Colorectal cancer is a very curable cancer if caught in early stages. Relative 5-year survival percentage for localized cases is 90.2%.

Classification models predicting different survival periods for example 1-year, 2-year, 3-year etc. survivability is still treated as a yes/no classification problem because the model's target variable, survival months, is divided each time into a binary target of either having survived n-years/months or not survived n-years/months, rather than having the model run once on a target variable split into multiple time ranges. From my trial and error in running models, I experienced that treating the classification model as a multiclass classification problem, decreased the model's performance for multiple reasons. Mainly models such as logistic regression do not inherently support multiclass prediction and requires in-depth knowledge to tweak the model according to the needs of the dataset to fit the model. Also, the decrease of sample size of each target label if the survival months are split into multiple ranges leading to more false negatives.

Other than using meta-classifiers to increase model performance, an interesting method seen in one paper, uses a "cluster-then-predict" method which improved accuracy. K-means was first used the clustering method, but HDBSCAN was found to be a more powerful method. Different classification models such as Naïve Bayes, logistic regression and random forest were not found to be that different in terms of performance. In tweaking the model, the target variable was first binned into 6 ranges, then 3 to improve balance. Linear regression was also used, but proved to be highly inaccurate, especially as survival months increased. In the final model, the target variable was split into 2 classes, either having survived past 5 years or not survived past 5 years.

Literature Review

Silva A., Oliveira T., Neves J., and Novais P., Treating colon cancer survivability prediction as a classification problem. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 5(1):37–50, 2016.
Dataset was split into 5 subsets according to survival times of 1,2,3,4 and 5+ years post cancer diagnosis. Features were selected from a list of 18 chosen by a colon specialist and narrowed down to six using decision trees and forward selection using RapidMiner. Meta-classifiers used with classification models to create ensemble models were AdaBoost, Bagging, Bayesian Boosting, Stacking and Voting. Classification models used were k-NN, NB, DT and RF. There were 14 classification models for both 6 and 18 attributes. Stacking was consistently most accurate across all years. The stacking model used k-NN, Decision Tree, and Random Forest classifiers as base learners, and a Naive Bayes classifier as a Stacking model learner. Decision tree was also a very effective model.

Li S., Razzaghi T. Personalized Colorectal Cancer Survivability Prediction with Machine Learning Methods. 2554-2558. 10.1109/BigData.2018.8622121, 2018.
Dataset was split into white, Hispanic and mixed dataset. LogR, RF, NN models were used and AdaBoost as an ensemble model. Random sampling and cost-sensitive learning was used to deal with unbalanced data. Models were used to predict 2-year survivability. Magnitude of coefficient for LogR and Gini impurity and entropy scores for RF and AdaBoost were used to select important features. Rank of important features were different for single races and mixed data subsets. Each predictor used on the single race subsets were more accurate than the mixed dataset.

Shukla N, Hagenbuchner M, Win KT, Yang J. Breast cancer data analysis for survivability studies and prediction. Comput Methods Programs Biomed. 2018 Mar; 155:199-208. doi: 10.1016/j.cmpb.2017.12.011. Epub 2017 Dec 12. PMID: 29512500.

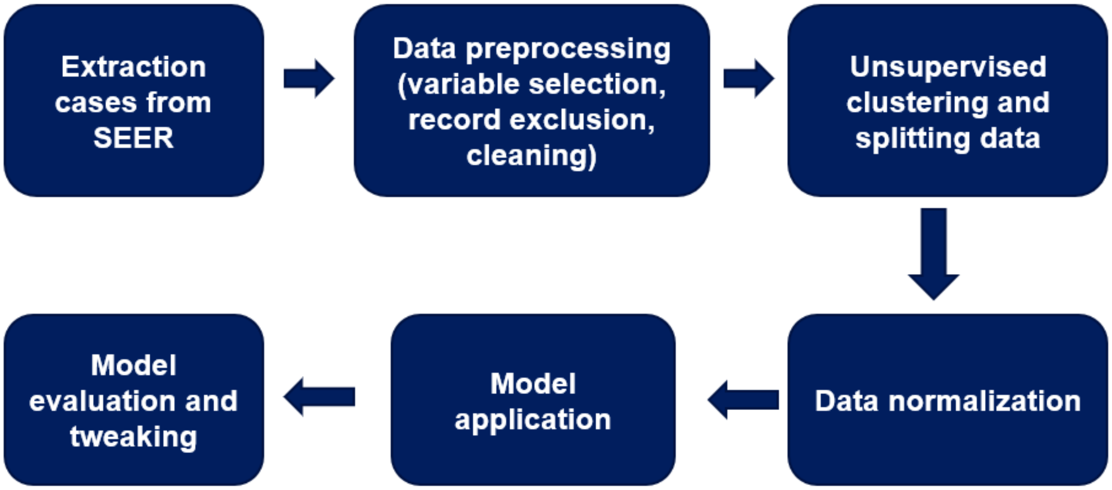
Paper predicts survivability in months for breast cancer. Unsupervised data mining method of self-organising map (SOM) to separate patients into groups with similarities and then density-based spatial clustering of applications with noise (DBSCAN) is used to extract and choose 9 groups. Information gain measured in entropy was used to choose 26 variables. Multilayer perceptron (MLP) model was applied to each cluster and accuracy was measured for the models' ability to predict 3, 5 and 7 year survivability. As survivability length increased, the models were less accurate, probably because the mean survivability months for each cluster were mostly between 30 - 65 months (2.5 to 5.5 years).

Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics. 2020 Dec;148:104371. DOI: 10.1016/j.ijmedinf.2020.104371.
3 deep learning models, ANN, CNN and RNN, compared to traditional machine learning models, NB, RF and SVM, were used to predict survivability in 3 classes (<=6 months, 6 months to 2 years, >2 years) in lung cancer. ANN, CNN, RNN again, as well as, Linear Regression, Gradient Boosting Machine and Random Forest Regression were used to predict survivability in months. RMSE increased as the survival period increased for regression models. Above 60 months, RMSE increased over 20%. For classification models, models had difficulty differentiating between the classes >=6 months and 6-2years. For target class (survivability), most of the data points were before 6-12 months, with a sharp decrease in data points when month range increased.

Al-Bahrani R., Agrawal A., Choudhary A. Colon cancer survival prediction using ensemble data mining on SEER data. 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 2013, pp. 9-16, doi: 10.1109/BigData.2013.6691752.
20 different classification schemes (various DTs, LogR, RF) along with meta-classifiers (bagging, adaboost, voting) are used to predict survival of colon cancer for target values (1 yr, 2 yrs, 5 yrs). 13 features were identified using correlation-based feature selection and information gain within Weka. The ensemble voting model was consistently the best performing model for all target classes. SMOTE was used to balance classes and this improved the model's output.

Li F., Duan Y. (2016) An Analysis of the Survivability in SEER Breast Cancer Data Using Association Rule Mining. In: Wang G., Ray I., Alcaraz Calero J., Thampi S. (eds) Security, Privacy and Anonymity in Computation, Communication and Storage. SpaCCS 2016. Lecture Notes in Computer Science, vol 10067. Springer, Cham. https://doi.org/10.1007/978-3-319-49145-5_19
Apriori algorithm was applied to 17 features of breast cancer data in SEER database. The data was first separated into "survived" and "not survived" categories. There were 12 nominal variables and 5 numeric variables chosen using a reference to a past paper. The numeric attributes are segmented and given numbers. Records that represent unknown / unmeasured are deleted. A random sample of the "survived" category is chosen to balance with the amount of data points in the "not survived" category. With support set at 20% and confidence set at 70%, 326 association rules for the "survived" category and 22 association rules for "not survived" were found. An issue observed with the Apriori algorithm is inefficiency because it needs to scan the entire database each time it calculates a new rule / itemset.

Flow Chart


Extraction cases from SEER
Data preprocessing (variable selection, record exclusion, cleaning)
Unsupervised clustering and splitting data
Data normalization
Model application
Model evaluation and tweaking

Preprocessing and EDA

Initial extract of data from SEER database for only Colon and Rectal cancers gives a dataset with 690,450 rows and 196 columns. After cleaning up columns and rows, final dataset to be used for modeling has 178,455 rows, 25 columns and 4 target variable columns.

The target variable is survival months, one column is continuous (in months), 3 are binned into different ranges. The ranges are 0-11, 12-23, 24-35, 36-47, 48-59 and 60+ months, 0-29, 30-59 and 60+ months, and 0-59, 60+ months.

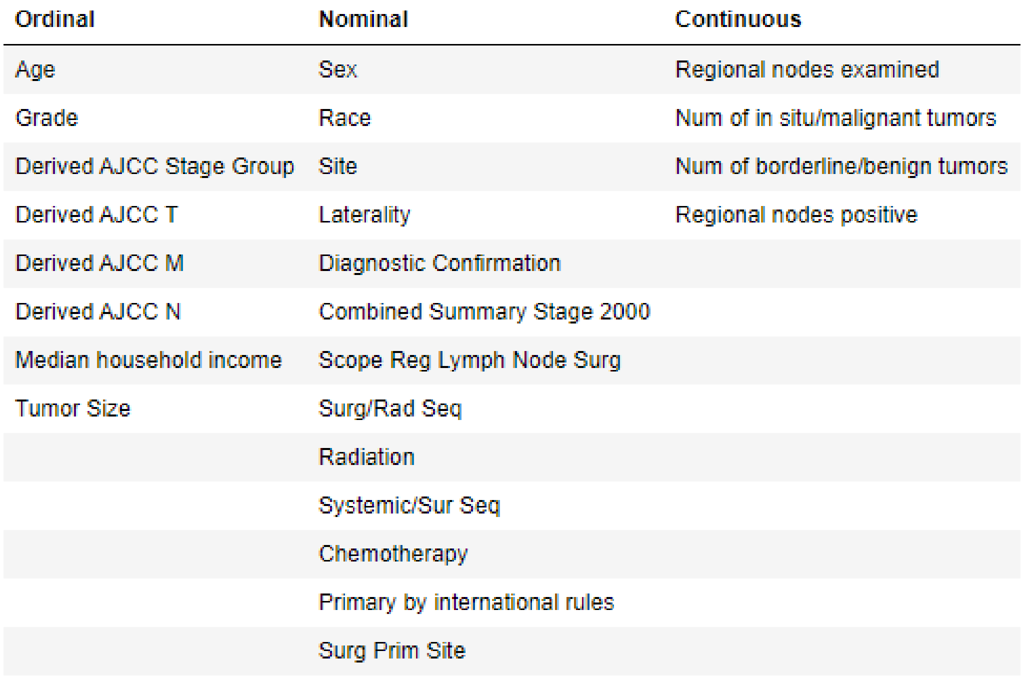
Methodology

A large portion of the project was spent on cleaning up the dataset as there were many variables and lengthy descriptions for each variable. I spent a lot of time pouring over the codes for each variable, trying to clean the data as much as possible.

After cleaning the data, data was split into train and test datasets, using 70% as train, 30% as test. Training dataset was clustered using K-means initially, test dataset was fit onto the train dataset, as to not disturb the clusters. Elbow method was used to determine clusters. Data was scaled to normalize the datapoints. Clusters were split into 0 or 1 dataset. Models were run on base dataset, 0 and 1 dataset. Models train and predicted on a specific target variable.

Initially the target variable, “Survival months”, was binned into 6 ranges, but the models were not performing well, for reasons mentioned in the introduction. Next, the target variable was binned into 3 ranges, to improve class imbalance, but this was not the best option either. Regression was also used to predict target variable to a specific month, but RMSE was very high. The last model used target variable as simply a binary variable of having survived or not survived within a 5-year period. Lastly, rather than K-means clustering, HDBSCAN was used as another unsupervised clustering method to see if the model could be improved.

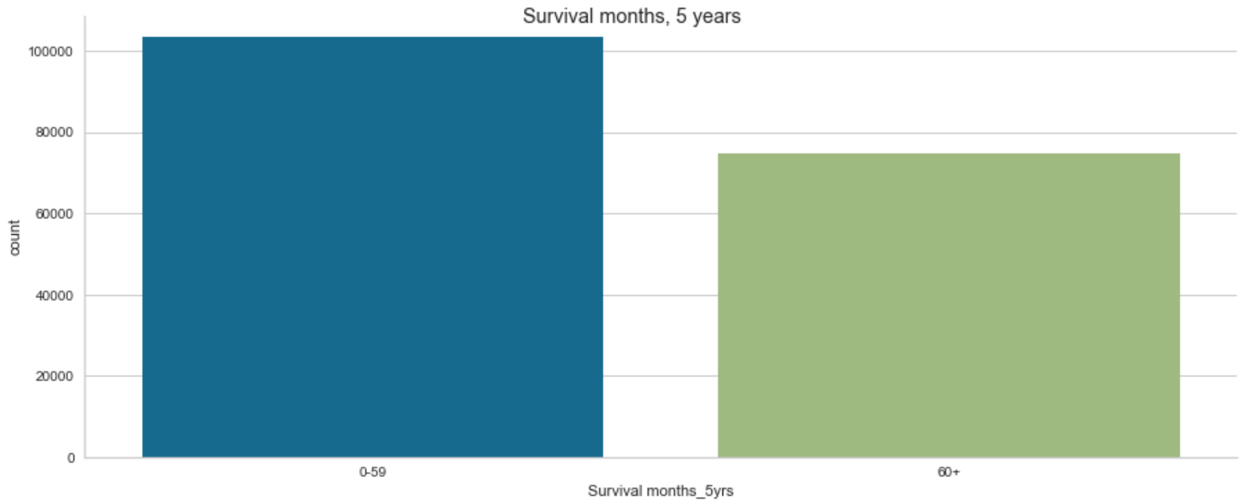
Dataset Description



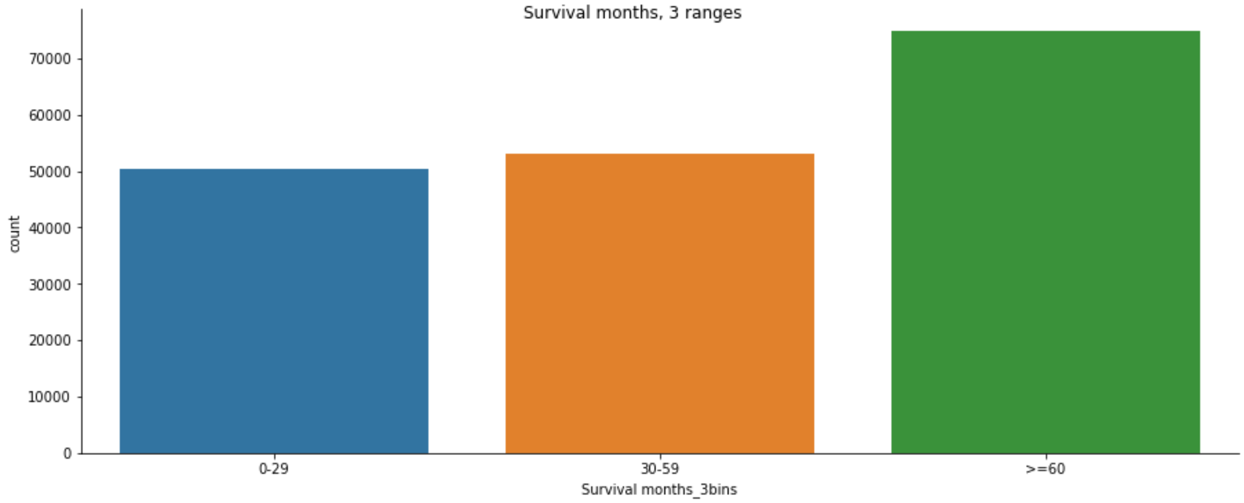
Ordinal	Nominal	Continuous
Age	Sex	Regional nodes examined
Grade	Race	Num of in situ/malignant tumors
Derived AJCC Stage Group	Site	Num of borderline/benign tumors
Derived AJCC T	Laterality	Regional nodes positive
Derived AJCC M	Diagnostic Confirmation	
Derived AJCC N	Combined Summary Stage 2000	
Median household income	Scope Reg Lymph Node Surg	
Tumor Size	Surg/Rad Seq	
	Radiation	
	Systemic/Sur Seq	
	Chemotherapy	
	Primary by international rules	
	Surg Prim Site	

The ordinal values and nominal values were label encoded using integers 1-k. For the ordinal values, the order was preserved as the integers were mapped onto the values according to the perceived order. This proved later to be not the best method, as it erroneously tells the model that there is a perceived order to the nominal values. As well, the distances between the categorical values cannot be "measured" or "labelled". A better method, if I had less variables, would be one-hot encoding.

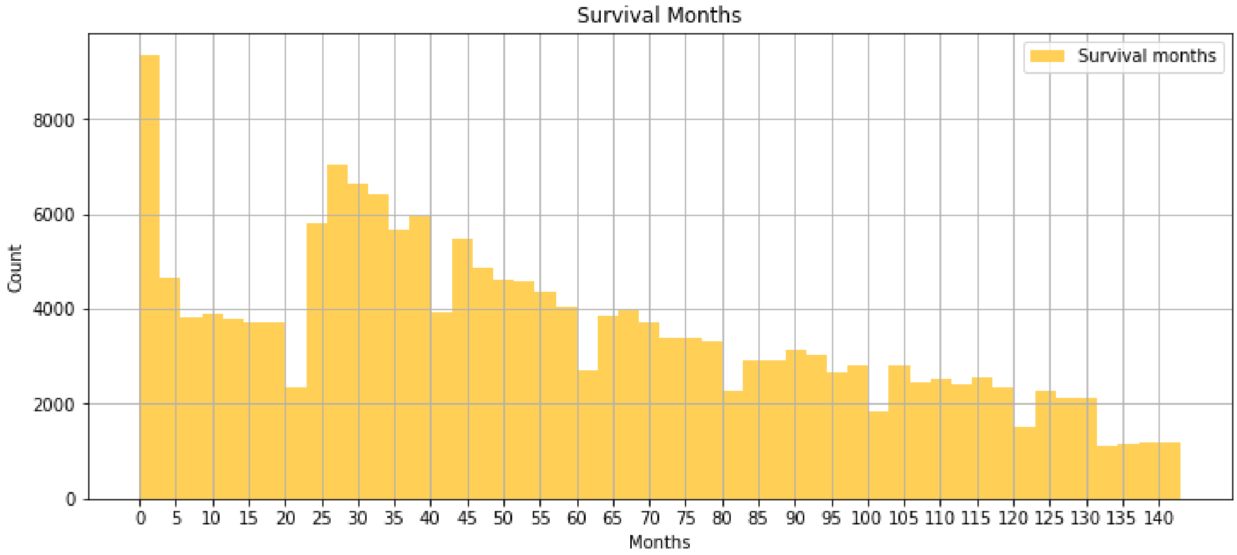
PCA was not used to reduce dimensionality because it does not work well with categorical values for the reasons mentioned above, as the "space" between the categorical values are meaningless. Multiple factor analysis may be a method I would look further into to reduce dimensionality.



Survival months split into less than 5 years (count: 103538) and 5 years or more (count: 74917).



Survival months split into 0-29 months, 30-59 months and 60 months or greater. (>=60: 74917, 30-59: 53096, 0-29: 50442)



Survival months as continuous variable, unit is months. There are peaks at 0 and 30 months. Overall right skewed, median is 50, slightly lower than mean which is 56.56.

Results

Clustering

Using elbow method, K-means returned 2 clusters. HDBSCAN, which does not assume clusters, but given minimum size of cluster greater than 500, gave 3 clusters. In HDBSCAN, cluster which is -1 is noise, and was removed.

Models

HDBSCAN, Logistic Regression – Target variable, 3 ranges (0-29 – Class 0, 30-59 – Class 1, 60+ months – Class 2)

Confusion Matrix for the Base Dataset

	Predicted 0	**Predicted 1**	**Predicted 2**
Actual 0	8224	1937	2930
Actual 1	4108	3928	6650
Actual 2	3409	4580	12995

Dataset	**F1-Score**	**Support**
Base	0.504	48761
Cluster 0	0.454	1900
Cluster 1	0.518	23410
Cluster 2	0.494	23451

HDBSCAN, Logistic Regression – Target variable, 2 ranges (0-59 – Class 0, 60+ months – Class 1)

Confusion Matrix for the Base Dataset

	Predicted 0	**Predicted 1**
Actual 0	15571	12206
Actual 1	5543	15441

Dataset	**F1-Score**	**Support**
Base	0.636	48761
Cluster 0	0.598	1900
Cluster 1	0.644	23410
Cluster 2	0.631	23451

K-means, Logistic Regression – Target variable, 3 ranges (0-29, 30-59, 60+ months)

Confusion Matrix for the Base Dataset

	Actual 0	**Actual 1**	**Actual 2**
Predicted 0	9542	2133	3353
Predicted 1	4382	4147	7536
Predicted 2	3467	4755	14222

Dataset	**F1-Score**	**Support**
Base	0.508	53537
Cluster 0	0.489	11697
Cluster 1	0.516	41840

K-means, Logistic Regression – Target variable, 2 ranges (0-59, 60+ months)

Confusion Matrix for the Base Dataset

	Predicted 0	**Predicted 1**
Actual 0	17482	13611
Actual 1	5659	16785

Dataset	**F1-Score**	**Support**
Base	0.641	53537
Cluster 0	0.622	11697
Cluster 1	0.645	41840

Linear Regression on Base Dataset

R^2 - 0.20495609122261438

RMSE - 33.776982044560825

Calculated Predictive Interval using Formula RMSE*2 = y+/- 67.5540

Analysis

Why did the initial multiclass classification models not work?

Unsurprisingly, the best model is the one where the target variable is binary. Unfortunately, most of my trial-and-error process was spent on a model which inherently would not work well as a multiclass prediction model. Even though the python library “sci-kit learn” does support a logistic regression model with multiclass target value, multiclass logistic regression complicates the original formula which is only supposed to predict probability between 0 or 1 (2 classes). Although ordinal logistic regression models could be used, this is something beyond my understanding and would take careful model building in python to ensure the proper model parameters are being used.

Why is K-means not the best clustering method?

K-means was a simple clustering method I was familiar with, although looking further, for my dataset was not ideal for K-means due to the high number of categorical features. K-means works better with continuous data as it measures Euclidian distances between datapoints to determine clusters. The sample space for categorical data is discrete and does noy have a natural origin. A Euclidean distance function on such a space is not very meaningful. Other assumptions that K-means make that is unsuitable to categorical data is that it assumes all clusters have centroids and same the shape, as well it assumes all variables have the same variance. Also, K-means is unable to handle noisy data and outliers.

Why HDBSCAN?

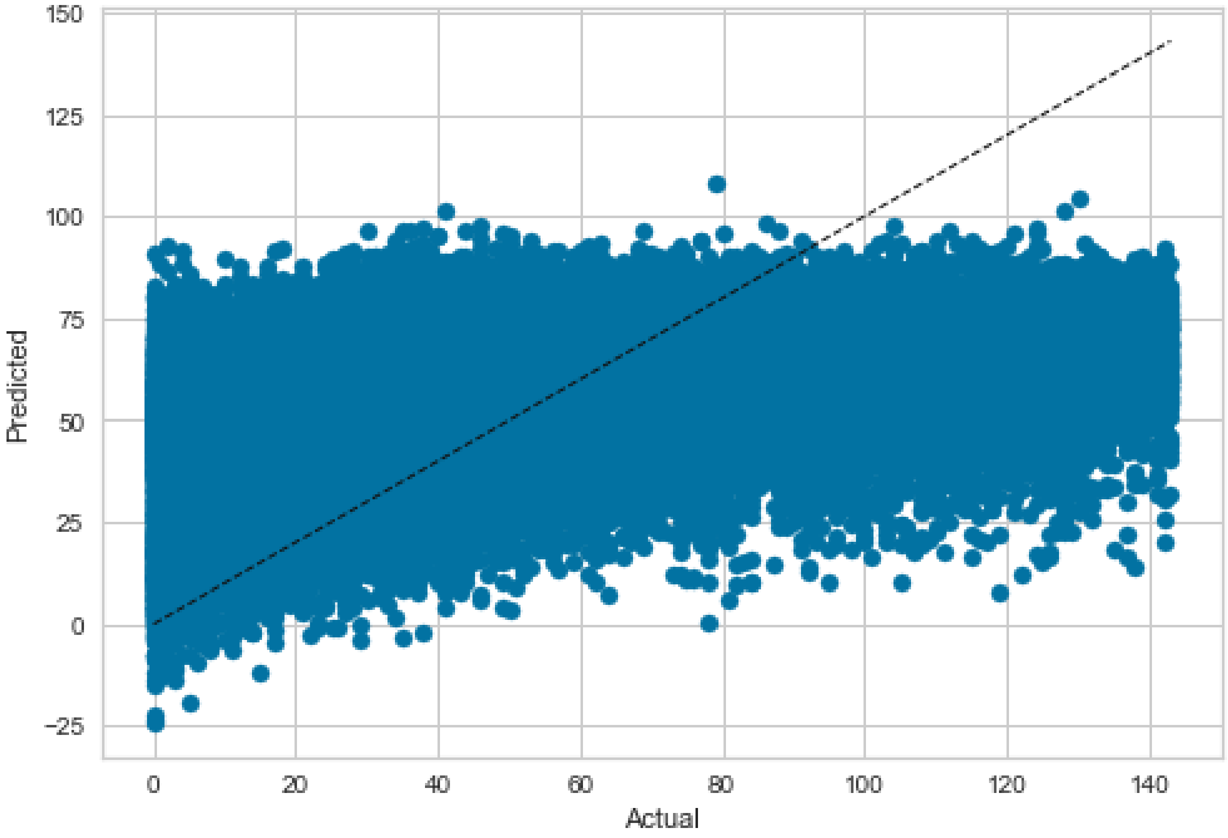
The paper I came across which used clustering prior to classification, used DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and SOM (Self-Organizing Maps) to create clusters. DBSCAN works better than K-means because it can detect outliers and detect irregularly shaped clusters. DBSCAN can cluster non-linearly separable clusters while K-means cannot. The paper I came across used a hybrid model of SOM and DBSCAN to improve clustering. SOM is good for reducing dimensionality which can improve DBSCAN. HDBSCAN is a improvement on DBSCAN, supposedly more powerful.

Problems with the final binary class model

For the binary class model results, from the confusion matrix, a big problem is false positives for class 1 (survived for 5 years). This is a big problem for this model and this research problem because in the case of prediction of survivability (class 1), it would be preferrable if there was an error, to have a higher false negative rate than false positive for class 1. False positives for class 1 indicate that the model mistakenly predicts someone will survive past 5 years, when in fact, their illness progression may be more serious, and they will not survive past 5 years. For this problem, this is a worse error than having false negatives for class 1, as in falsely predicting someone NOT survive past 5 years, when they will.

Linear Regression Model

The linear regression model, we see actual vs. predicted values start to deviate as survival months increases. At months greater than around 90, the model always predicts a lower survival range. Even though the model can predict the correct ranges in the earlier months, there is a very large prediction interval, especially in the earlier survival months. The calculated prediction range from RMSE is +/- 67.55, which renders this model useless as a classification model with better performance would be more useful.



Future Considerations

Starting with the model with the binary target variable, I would find a method of feature selection that filters the variables so that it provides a better performing model. I used chi square to compare variables when the target variable was categorical to find the variables that were most correlated, but the filtered dataset was not an improvement.

Once the dataset is filtered to include less columns, rather then labelling each categorical column with 1-k integers, I would use one hot encoding to transform each variable. Currently with 25 variables, I did not want to increase the dimensionality of the dataset so drastically, as a lot my categorical variables had multiple values which would result in many more columns. One hot encoding would ensure the model does not assume levels within the categorical variables.

For HDBSCAN clusters, increasing the minimum size of each cluster would improve the model, as the smallest cluster in HDBSCAN is too small which increases sampling variability and higher false negatives.

Logistic regression was used as it was a simple model that I had already come across, but I would try to use a neural network to model the data, although a more complex model does not mean better results, neural networks were frequently used in past papers with promising results. I would try this one I am more familiar with how to apply such a model.

Voting ensembles are also frequently seen in literature, so I would attempt to apply voting ensemble which can sum the predictions made by several classification models.

To improve upon the dataset, I would gather non-clinical information such as family history and lifestyle choices which have a big impact on one developing colon cancer. In terms of clinical data, if genetic testing was available, then having a variable that indicates presence if genetic markers for this cancer would be helpful as genetic testing is very accurate.

References

1) Lynch CM, Abdollahi B, Fuqua JD, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform*. 2017;108:1-8. doi:10.1016/j.ijmedinf.2017.09.013

2) Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches [published online ahead of print, 2020 Dec 29]. *Int J Med Inform*. 2020;148:104371. doi:10.1016/j.ijmedinf.2020.104371

3) Hegselmann, S., Gruelich, L., Varghese, J. &amp; Dugas, M.. (2018). Reproducible Survival Prediction with SEER Cancer Data. *Proceedings of the 3rd Machine Learning for Healthcare Conference, in PMLR* 85:49-66

4) Li, S., & Razzaghi, T. (2018). Personalized Colorectal Cancer Survivability Prediction with Machine Learning Methods*. *2018 IEEE International Conference on Big Data (Big Data)*, 2554-2558.

5) Delen, Dursun. (2009). Analysis of Cancer Data: A Data Mining Approach. Expert Systems. 26. 100-112. 10.1111/j.1468-0394.2008.00480.x.

6) Bartholomai JA, Frieboes HB. Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *Proc IEEE Int Symp Signal Proc Inf Tech*. 2018;2018:632-637. doi:10.1109/ISSPIT.2018.8642753

7) Cooper GS, Virnig B, Klabunde CN, Schussler N, Freeman J, Warren JL. Use of SEER-Medicare data for measuring cancer surgery. *Med Care*. 2002;40(8 Suppl):IV-48. doi:10.1097/00005650-200208001-00006

8) Senders JT, Staples P, Mehrtash A, et al. An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery*. 2020;86(2):E184-E192. doi:10.1093/neuros/nyz403

9) Lynch CM, van Berkel VH, Frieboes HB. Application of unsupervised analysis techniques to lung cancer patient data. *PLoS One*. 2017;12(9):e0184370. Published 2017 Sep 14. doi:10.1371/journal.pone.0184370

10) Shukla N, Hagenbuchner M, Win KT, Yang J. Breast cancer data analysis for survivability studies and prediction. *Comput Methods Programs Biomed*. 2018;155:199-208. doi:10.1016/j.cmpb.2017.12.011

11) R. Al-Bahrani, A. Agrawal and A. Choudhary, "Colon cancer survival prediction using ensemble data mining on SEER data," *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, 2013, pp. 9-16, doi: 10.1109/BigData.2013.6691752.

12) Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*. 2005;34(2):113-127. doi:10.1016/j.artmed.2004.07.002

13) Silva, Ana & Oliveira, Tiago & Neves, João & Novais, Paulo. (2016). Treating Colon Cancer Survivability Prediction as a Classification Problem. *Adcaij: Advances In Distributed Computing And Artificial Intelligence Journal*. 5. 37. 10.14201/ADCAIJ2016513750.

14) Gao, Peng & Zhou, Xin & Wang, Zhen-Ning & Song, Yong-xi & Tong, Lin-lin & Xu, Yingying & Yue, Zhen-yu & Xu, Hui-mian. (2012). Which Is a More Accurate Predictor in Colorectal Survival Analysis? Nine Data Mining Algorithms vs. the TNM Staging System*. PloS one*. 7. e42015. 10.1371/journal.pone.0042015.

15) Mourad, Moustafa & Moubayed, Sami & Dezube, Aaron & Mourad, Youssef & Park, Kyle & Torreblanca-Zanca, Albertina & Torrecilla, Jose & Cancilla, John & Wang, Jiwu. (2020). Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. *Scientific Reports*. 10. 10.1038/s41598-020-62023-w.

16) Silva A., Oliveira T., Novais P., Neves J., Leão P. *Advances in Intelligent Systems and Computing.* Volume 476. Springer; Berlin, Germany: 2016. Developing an individualized survival prediction model for colon cancer; pp. 87–95. Chapter Developing.

17) Ryu SM, Lee SH, Kim ES, Eoh W. Predicting Survival of Patients with Spinal Ependymoma Using Machine Learning Algorithms with the SEER Database [published online ahead of print, 2018 Dec 28]. *World Neurosurg*. 2018;S1878-8750(18)32914-0. doi:10.1016/j.wneu.2018.12.091

18) Song Y, Gao S, Tan W, Qiu Z, Zhou H, Zhao Y. Multiple Machine Learnings Revealed Similar Predictive Accuracy for Prognosis of PNETs from the Surveillance, Epidemiology, and End Result Database. *J Cancer*. 2018;9(21):3971-3978. Published 2018 Oct 10. doi:10.7150/jca.26649