CS3244 Team 30 List of References

EDA and Data Preprocessing

Pramoditha, R. (2021, December 16). Encoding Categorical Variables: One-hot vs Dummy Encoding. *Towards Data Science*. https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db

usεr11852. (2018, April). *Featurization before or after dataset splitting [Online forum post]*. StackExchange. https://stats.stackexchange.com/questions/338400/featurization-before-or-after-dataset-splitting

Ander Biguri. (2015, June). *Significance of 99% of variance covered by the first component in PCA [Online forum post]*. Stackoverflow. https://stackoverflow.com/questions/30777569/significance-of-99-of-variance-covered-by-the-first-component-in-pca

Omamalin, S. J. (2019). *Multicollinearity and how it affects your model*. Kaggle. https://www.kaggle.com/code/sjodcre/multicollinearity-and-how-it-affects-your-model

Allison, P. (2012, September 10). *When Can You Safely Ignore Multicollinearity?* Statistical Horizons. https://statisticalhorizons.com/multicollinearity/

Jermain, N. (2019, June 24). *Transforming Skewed Data for Machine Learning*. Open Data Science. https://opendatascience.com/transforming-skewed-data-for-machine-learning/

-. S. (2020). *Data Dictionary*. Kaggle. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/discussion/119320?datasetId=426827

Brownlee, J. (2020, August 20). *How to Use StandardScaler and MinMaxScaler Transforms in Python*. Machine Learning Mastery. https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/?utm_source=pocket_saves

Muaz, U. (2019, July 25). *Autoencoders vs PCA: when to use ?* Medium. https://towardsdatascience.com/autoencoders-vs-pca-when-to-use-which-73de063f5d7

Wang, Z. (2018, August 10). *Practical tips for class imbalance in binary classification*. Medium. https://towardsdatascience.com/practical-tips-for-class-imbalance-in-binary-classification-6ee29bcdb8a7

Mangale, S. (2020, August 28). *Scree Plot*. Medium. https://sanchitamangale12.medium.com/scree-plot-733ed72c8608

jamesmf. (n.d.). *Why does applying PCA on targets causes underfitting? [Online forum post]*. StackExchange. https://datascience.stackexchange.com/questions/8087/why-does-applying-pca-on-targets-causes-underfitting

Baretto, P. (2020, June 3). *Removing Multicollinearity for Linear and Logistic Regression.* Medium.

https://medium.com/analytics-vidhya/removing-multi-collinearity-for-linear-and-logistic-regression-f1fa744f3666

Is standardization needed before fitting logistic regression? (n.d.). Cross Validated. Retrieved November 15, 2022, from https://stats.stackexchange.com/questions/48360/is-standardization-needed-before-fitting-logistic-regression

Pulagam, S. (2020, August 1). *Feature Scaling — Effectively Choose Input Variables Based on Distributions*. Medium. https://towardsdatascience.com/feature-scaling-effectively-choose-input-variables-based-on-distributions-3032207c921f

Lee, W.-M. (2022, February 2). *Using Principal Component Analysis (PCA) for Machine Learning*. Medium. https://towardsdatascience.com/using-principal-component-analysis-pca-for-machine-learning-b6e803f5bf1e

Labelling

Seanny. “Credit Card Approval Prediction Using ML.” *Kaggle.com*, 2020, www.kaggle.com/code/rikdifos/credit-card-approval-prediction-using-ml. Accessed 15 Nov. 2022.

SMOTE/ SMOTENC

Brownlee, J. (2020, January 16). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Keller, J. (2020, January 30). *Upsampling with SMOTE for Classification Projects*. Medium. https://towardsdatascience.com/upsampling-with-smote-for-classification-projects-e91d7c44e4bf

SATPATHY, S. (2020, October 6). *SMOTE - A Common Technique to Overcome Class Imbalance Problem*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Wijaya, C. Y. (2021, October 12). *5 SMOTE Techniques for Oversampling your Imbalance Data*. Medium. https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5

Lema, G., Nogueira, F., & Aridas, C. K. (n.d.). *SMOTENC — Version 0.9.1*. Imbalanced Learn. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html

Logistic Regression

Brownlee, J. (2016, September 22). *Logistic Regression for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Brownlee, J. (2020, January 27). *Cost-Sensitive Logistic Regression for Imbalanced Classification*. Machine Learning Mastery.

https://machinelearningmastery.com/cost-sensitive-logistic-regression/

jazib jamil. (2015, March). Controlling the threshold in Logistic Regression in Scikit Learn [Online forum post]. Stackoverflow. https://stackoverflow.com/questions/28716241/controlling-the-threshold-in-logistic-regression-in-scikit-learn

M, D. (2022, July 10). Handling imbalanced data with class weights in logistic regression. *Analytics India Magazine*. https://analyticsindiamag.com/handling-imbalanced-data-with-class-weights-in-logistic-regression/

Andreas Mueller. (2015, June). *How does the class_weight parameter in scikit-learn work? [Online forum post]*. Stackoverflow. https://stackoverflow.com/questions/30972029/how-does-the-class-weight-parameter-in-scikit-learn-work

Jermain, N. (2019, June 6). *Strategies for Addressing Class Imbalance*. Open Data Science. https://opendatascience.com/strategies-for-addressing-class-imbalance/

Dino, L. (2022, April 23). *Define threshold of logistic regression in Python*. Medium. https://medium.com/@24littledino/define-threshold-of-logistic-regression-in-python-56c60664fc3e

rohan007. (n.d.). *Stratified K Fold Cross Validation*. GeeksforGeeks. https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/?utm_source=pocket_saves

Pedregosa, F. et al (n.d.). *sklearn.linear_model.LogisticRegression*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Decision Tree

scikit learn. (2009). *1.10. Decision Trees — scikit-learn 0.22 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/tree.html

What Is a Decision Tree and How Is It Used? (n.d.). Careerfoundry.com. https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/#:~:text=Decision%20trees%20are%20extremely%20useful

Random Forest

Koehrsen, Will. “Random Forest in Python.” *Medium*, Towards Data Science, 27 Dec. 2017, towardsdatascience.com/random-forest-in-python-24d0893d51c0.

Naviani, Avinash. “Sklearn Random Forest Classifiers in Python Tutorial.” *Www.datacamp.com*, May 2018, www.datacamp.com/tutorial/random-forests-classifier-python.

“Credit Card Predictive Analysis.” *Kaggle.com*, 2020, www.kaggle.com/code/umerkk12/credit-card-predictive-analysis. Accessed 15 Nov. 2022.

AdaBoost

Akash Desarda. (2019, January 17). *Understanding AdaBoost*. Medium; Towards Data Science. https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe

Brownlee, J. (2020, April 30). *How to Develop an AdaBoost Ensemble in Python*. Machine Learning Mastery. https://machinelearningmastery.com/adaboost-ensemble-in-python/

Delacruz, C. (2021, July 2). *How To Include An ADA Boost Model's Base Estimator In A Grid Search When It's Contained In A…*. Medium. https://c-delacruz.medium.com/how-to-include-an-ada-boost-models-base-estimator-in-a-grid-search-when-it-s-contained-in-a-b328568a2b83

python - Using GridSearchCV with AdaBoost and DecisionTreeClassifier. (n.d.). Stack Overflow. Retrieved November 15, 2022, from https://stackoverflow.com/questions/32210569/using-gridsearchcv-with-adaboost-and-decisiontreeclassifier

Starmer, J. (2019). AdaBoost, Clearly Explained [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=LsK-xG1cLYA

XGBoost

Amy. (2021, November 7). *Hyperparameter Tuning For XGBoost: Grid Search Vs Random Search Vs Bayesian Optimization*. Grab N Go Info. https://grabngoinfo.com/hyperparameter-tuning-for-xgboost-grid-search-vs-random-search-vs-bayesian-optimization/

Analytics Vidhya. (2016, March). *Complete Guide to Parameter Tuning in XGBoost (with codes in Python)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

Okamura, S. (2020, December 30). *GridSearchCV for Beginners*. Medium. https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee

Brownlee, J. (2020, February 5). *How to Configure XGBoost for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-for-imbalanced-classification/

Clustering

sklearn.cluster.KMeans — scikit-learn 0.21.3 documentation. (2019). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Kumar, Ch. N. S., Rao, K. N., Govardhan, A., & Sandhya, N. (2015). Subset K-Means Approach for Handling

Imbalanced-Distributed Data. *Advances in Intelligent Systems and Computing*, 497–508. https://doi.org/10.1007/978-3-319-13731-5_54

LIME

Wijesinghe, Vikum. “Explaining Random Forest Model with LIME.” *Kaggle.com*, 2019, www.kaggle.com/code/vikumsw/explaining-random-forest-model-with-lime. Accessed 15 Nov. 2022.

Banerjee, P. (2019). *Explain your model predictions with LIME*. Kaggle. https://www.kaggle.com/code/prashant111/explain-your-model-predictions-with-lime/notebook

Kuo, C. (2020, February 20). *Explain Your Model with LIME*. Medium. https://medium.com/dataman-in-ai/explain-your-model-with-lime-5a1a5867b423

Statement of Independent Work

University of Pretoria. (n.d.). Assignment Front Page and Declaration. In *University of Pretoria*. https://www.up.ac.za/media/shared/25/Forms/assignment_mark_frontpage.docx