Jeffery Dirdem
ITAI 1371
December 8, 2024

Final Project Report

I would like to start with the fact that I was not able to find the file suggested to complete this project how it was originally intended because I actually cleared all files from my hard drive for computational performance issues so I took it upon myself to create my own. The goal of this study was to compare the effectiveness of two machine learning methods, Lasso Regression and Support Vector Machine (SVM) Regression, in maximizing profit margins. The models' accuracy and fit were evaluated using performance indicators like RMSE and R-squared (R^2). The investigation was carried out on a synthetic dataset intended to replicate real-world settings. To replicate a realistic scenario, a dataset of 1000 samples and 10 characteristics was created. Each feature was randomly assigned a value ranging from 0 to 100. The target variable, which represents profit margins, was created as a linear combination of selected parameters with additional noise to account for unpredictability. This dataset was turned into a pandas DataFrame to make processing easier.

The data preparation procedure included isolating the characteristics from the target variable. The dataset was then separated into training and testing subsets, with 80% used for training and 20% for testing. Numerical scaling was used to maintain uniformity across feature magnitudes, which is critical for improving the performance of both the Lasso and SVM models. Lasso Regression was chosen due to its capacity to manage feature selection using L1 regularization. The model was cross-validated five times to determine the best regularization

strength. SVM Regression, which is known for its ability to model complicated, non-linear connections, was implemented with the radial basis function (RBF) kernel. GridSearchCV was used to tune SVM hyperparameters by iteratively testing combinations of the regularization parameter ("C") and kernel coefficient (“gamma”). After evaluation, the Lasso Regression model had an RMSE of 4.90 and a R^2 value of 0.90. These findings suggested that the model correctly caught a considerable percentage of the variance in profit margin data, albeit with some degree of error. The optimized SVM Regression model had an RMSE of 4.37 and a R^2 value of 0.92. This greater performance revealed that SVM Regression better captured the dataset's complicated relationships. The findings emphasized the need of scaling numerical features and adjusting hyperparameters in machine learning models. Lasso Regression's strength is its ability to identify and accentuate the most important features, making it appropriate for circumstances where interpretability is vital. SVM Regression revealed a stronger capacity to fit the data appropriately, which is useful for prediction problems with non-linear dependencies.

In conclusion, SVM Regression was chosen as the best model for this research because to its greater accuracy and fit. Future research could build on this work by comparing these algorithms to more complex models like Random Forests or Neural Networks. This would assist in determining the best effective method for estimating profit margins in various circumstances. Furthermore, methods for interpreting complicated models, such as feature importance analysis or SHAP values, may provide useful insights for decision-making processes.

Bibliography

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

2. McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56.

3. Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering, 13*(2), 22-30.

4. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

5. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297.