Using Feature Engineering Techniques to Improve Multivariate Model Forecasting of the South African Unemployment Rate

Introduction.

What is feature engineering?

- Feature Engineering is a machine learning process for extracting useful features out of raw data using domain knowledge.
- It aims to find significant features, through processes such as combining relatively insignificant features to contribute to the model.

Why a paper on feature engineering?

- Previous work on improving the performance of multivariate machine learning forecasting models focused on feature selection techniques.
- They often do not cover the entire space of feature Optimization, limiting the scope of the problem.
- Using feature engineering to improve the performance of the models each feature has the potential to be engineered to have a meaningful impact on the final model.
- Every feature in the data set is seen as useful rather than selecting only a subset and leaving other features unexplored.

Why we chose to model the south African unemployment rate?

- South Africa's unemployment rate shows no sign of slowing down, in the fourth quarter of 2020 the rate was 32.5% the highest ever recorded
- This raises concerns regarding the well-being of the country. The South African market is still not creating enough jobs as more people enter the labour market.
- Accurately predicting the unemployment rate can help policymakers deal with the unemployment crisis.

Methods.

Data.

- The Data used in this paper was sourced from the South African Reserve Bank.
- SVR was used as the model for testing various techniques
- The process of engineering features can be summarised into four categories.

Feature Engineering Techniques.

- **Lags** This process explores the impact of adding lags to the data. The focus is on how many lags are best suited to improve the models forecasting ability. For this experiment, two lags were added to the data.
- **Statistical Features** These features were sourced using statistical measures. For each feature in our base dataset.
- **Fast-Fourier transform (FFT)** The Fourier transform is a function that takes time-domain signals into the frequency domain. These features were created by taking the Fourier transform of the base dataset and applying statistical transformations to obtain a total of 611 features.
- **tsFresh features** Before these features were extracted a correlation between the original 47 features and the target variable revealed the most highly correlated feature.

- The total number of features that were engineering ended up at 3237.
- Model features selection was applied with the model of choice being *linear regression*.

- This reduced the number of new features to a value of 359.

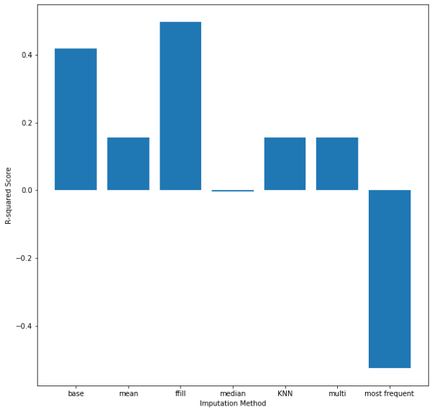
Results and Analysis

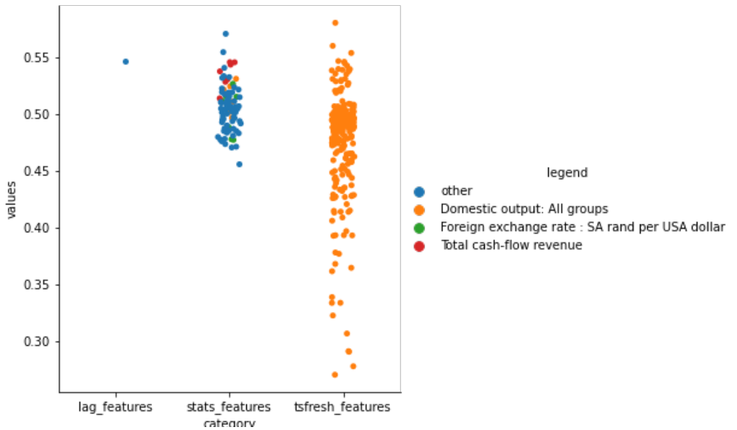
Feature Imputation.

- The first feature engineering technique that was applied in this paper was imputation.
- five imputation methods namely Forward fill, mean imputation, median imputation, most frequent imputation, multivariate imputation and KNN imputation were applied to the original data.
- Forward fill imputation gave the best R-squared score of 0.496.
- The lowest forward fill score recorded is -0.61, and the lowest score overall was -0.67 and comes from Multivariate imputation.
- The highest score obtained by Multivariate imputation is - 0.015.

Feature Extraction.

- Following imputation techniques, experiments were run on the 359 features that were extracted across the different categories.
- Each feature was added to the base data set one at a time and an SVR model was run on it after the addition of the feature.
- Across all four categories of selected features, the tsFresh category had 272 features out of 359 features.
- The average R-squared score for this category is 0.547083.
- The feature that improved the base data set the best also came from this category with an R-squared score of 0.580946.
- The worst performing category is Fourier Transform Features. No features from this category were selected.
- The lowest R-squared score recorded is 0.477384 this is lower than our base dataset score with no added features.





UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG