

EXPERIMENT REPORT

| | |
|--------------|--------------------------------------|
| Student Name | Nathan Collins |
| Project Name | MLAA_Assignment_1 |
| Date | |
| Deliverables | <Part A Report> <Part A Notebook> |

1. EXPERIMENT BACKGROUND

Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.

1.a. Business Objective

Explain clearly what is the goal of this project for the business.

The goal of the project is to develop a model that can **accurately** predict cancer mortality in different US counties based on a 33-feature dataset.

How will the results be used?

The insights provided can be used to improve cancer prevention, diagnosis, and treatment strategies, ultimately reducing the cancer mortality rate in the US.

These results may come in the form of a myriad of benefits, such as aiding local healthcare employees and providers to retain sufficient inventory to assist cancer patient health outcomes (financial, age, gender), assisting leaders and policymakers in providing adequate funding to US counties of higher cancer incidence (geography, death rate), and assisting decisions to investigate higher-risk areas and fund investigations to develop interventions (government and private health cover).

Results may also assist citizens in comprehending a family member's survivability following diagnosis. Constructing a successful model ultimately reduces the prevalence of cancer or increases patient survival by identifying linked features.

What will be the impact of accurate or incorrect results?

The influence of accurate results could affect how cancer mortalities are supplied for, intervened and prevented. As we are

dealing with lives and collateral that comes with loss, the stakes are high should incorrect results be generated and utilised. An example may include a model that identifies a county with a high cancer mortality rate, healthcare providers can allocate increased cancer screening and prevention efforts in that county.

Likewise, if the generated results prove unreliable, it may lead to a misallocation of crucial resources, resulting in ineffective interventions and more fatalities than necessary. Additionally, should a successful model be constructed, mispredictions on survivability may leave lasting emotional and financial collateral to the diagnosed and their peripheral family.

Present the hypothesis you want to test, the question you want to answer or the insight you are seeking.

Part C hypothesis:

Feature engineering data features will provide more accurate Mean Squared Errors, than without feature engineering.

A Random Forest Regression model will provide the most accurate Mean Squared Error for engineered features, than other Multivariate Regression models.

Explain the reasons why you think it is worthwhile considering it.

The hypothesis warrants exploration, as it engages with and applies a variety of multiple regression models against a data frame that is yet to provide an accurate MSE evaluation.

The hypothesis also explores depth behind the data features through feature engineering, a tool that may reveal further relationships between variables that simple correlations and previous regression models can't highlight.

Additionally, when consulting the theory, the hypothesis indirectly infers there is causality between certain demographic and socioeconomic factors and cancer mortality rates in US counties.

Should these relationships be confirmed, it serves to aid social awareness and acknowledgement, leading to ample provision of public health policies and interventions aimed at reducing cancer mortality rates in areas with these correlating features. This has the overall affect at saving human life, a priceless faculty to harness. Outcomes from the following may furthermore serve as a benchmark for current health standards, facilitating an understanding whether conditions improve over future generations.

1.b.
Hypothesis

1.c.
Experiment
Objective

Detail what will be the expected outcome of the experiment.

The expected outcome for **Part C**, is the MSE evaluations of a each regression model instantiated, will providing different degrees of accuracy, each a representative of a different perspective with viewing the data frame. The feature engineering will augment this further and provide another angle that will enhance accuracy and lower previous MSE and MSA scores. As the topic of exploration requires more features to entirely visualise, these evaluation metrics will appear lower, though it isn't anticipated to be a finalised and noticeably accurate outcome.

List the possible scenarios resulting from this experiment.

It is likely no model is results in a successful prediction of a complex issue, like cancer. This outcome would entail all train and test models yield MSE and MSA's evaluations, unchanged, or still relatively higher from previous model iterations. Another highly probable extension of this, is a lack of noticeable correlation between variables that underwent feature engineering. Should this be the case, resolving the scenario would be to explore other features, or perform feature engineering to discern relationships, if present.

The other possible scenario is the outcome that meets the business goal, which is the creation and deployment of an accurate model following feature engineering and testing alternative regression methods. <As human behaviour is chaotic, variables will never truly be 100% defined, thus success is gauged deemed through minimal error with predictions. The outcome could have shape lives, as mentioned previously.>

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment.

List the different steps/techniques used and explain the rationale for choosing them.

Describe the steps taken for preparing the data (if any).

<No further data preparation was carried out following part A, aside from including the Geography feature and performing subsequent feature engineering on five columns. These are detailed below.>

List also the steps you decided to not execute and the reasoning behind it.

- Sort Geography by county. This would result in the majority of the variables returning "1". It is more effective and easier to visualise by state.
- Integrating the "SomeCollege18" features in higher tertiary education, as this feature contained an excess of NaN values.
- Removing some noticeable outliers, as they represented real-world scenarios (minus ages exceeding 300). While this may result in an overall reduced accuracy, it is a more realistic representation of the world.
- Removing the inserted averages for "PctPrivateCoverageAlone" and "PctEmployed16_Over". While not an accurate representation of the dataset, these features expressed higher correlations with the target variable, and were too difficult to exclude. It should be noted that integrating these features may result in a minor
- Inserting more features from external data frames. While trailed, this decision was revoked, as it would have different contextual circumstances regarding collection methods, requiring additional standardisation, and may shift the results to one that doesn't reflect real scenarios.

Highlight any step that may potentially be important for future experiments

"Reducing Geography features into simply States and converting Geography into numerical features through one-hot-encoding."

Should the MSE evaluations of each model provide accurate feedback and a successful model is deployed, discerning whether the features differ across the US states facilitates a targeted approach at reducing cancer mortality rates.

Explain the rationale why you had to perform these steps.

Feature Engineering:

Reducing “Geography” into State (instead of County, State), facilitated nominal-ordinal data conversion with the one-hot-encoder library. This step enables the capacity to sort features, by the area of origin.

Higher tertiary education: Concatenating bachelor’s degrees (18-24 and 25 and over) granted a pool of highly educated individuals. This data set would prove valuable, as it helps discern whether the top ends of education qualification also amounts to knowledge to pursue appropriate intervention if diagnosed. Additionally, more knowledge would correspond with safer conduct, which would also lead to fewer manifestations of cancer through lifestyle choices.

Non-Caucasian descent helps reaffirm existing statistics unearthed throughout the literature review, conducted prior to the investigation. It was stated that more ethnic backgrounds experience higher mortality rates due to cancer. By concatenating all groups identifying as non-white, it may help improve the accuracy of the model.

“Male-Female Median Age Gap”. As women statistically live longer than men, it is possible this also penetrates cancer mortality rates. Additionally, men and women experience gender specific cancers that are more apparent than others. As the median age of men and women were the only values included that differentiate the genders, (and following several dicey scripts attempting to derive further value) the least destructive, though still subtly contributes a new feature, is their difference in average ages. This followed previous cleaning of the age outliers that exceeded 300.

Discerning the percentage of the population that is uninsured (without public coverage alone, or private coverage alone), could also provide a valuable angle, as both insurances positively and negatively correlate with mortality rate. This was determined by subtracting 100 from the two columns, concatenated.

List also the feature you decided to remove and the reasoning behind it.

“binnedInc” was removed during pre-processing, as it was a non-integer feature and encapsulated another feature already available in the data, “Median Income”. The ID column was also removed, at it offered no descriptive features about the business objective.

Four other columns generated throughout the feature engineering phase were also cut due to their negligible effect in improving accuracy evaluations of finalised models. These included combining and visualising variables from external data frames (CO2 emissions, diet, weight) with median male/female life

expectancies, and visualising their effect on cancer mortality rates. These external datasets had misaligned dates (not between 2010-2016). Such ambiguity around the times of collection served as additional rationale to discard these features.

Highlight any feature that may potentially be important for future experiments.

“Reducing Geography features into simply States and converting Geography into numerical features through one-hot-encoding.”

Should the MSE evaluations of each model provide accurate feedback and a successful model is deployed, discerning whether the features differ across the US states facilitates a targeted approach at reducing cancer mortality rates.

Describe the model(s) trained for this experiment and why you choose them.

Lasso Regression was selected, as it uses L1 regularisation to select features. The benefit of shrinking coefficients of less important features to zero, may produce a simpler and more interpretable model, that may improve accuracy evaluations. This may prove valuable given the spread of the data is quite broad and irregular.

Ridge Regression was selected as it typically identifies multicollinearity and prevents overfitting. By adding a penalty term to the cost function, the magnitude of coefficients remain restricted, shaping a simpler and stable model. This technique was selected, as the differences between some variable distributions and values are more noticeable than others – namely the percentages of county residents with specified health coverages.

K-Nearest Neighbours Regression (KNN) was selected, as it is a non-parametric algorithm with a classification function. By training a model to identify a data point's neighbouring data (k) and using their labels or values to make a prediction, it may remedy the inconsistent sampling of the data that results in some distributions lacking linearity.

As Random Forest Regression constructs a multitude of decision trees, the combination power behind their predictions may improve accuracy of existing evaluation metrics. It also prevents overfitting. The random selection of features of this model, which ultimately lead to subsequent subsets of features for aggregation, may help coalesce the influence of the spread of the data frame (as witnessed in previous models). Random Forest is more defined in its decision tree synthesis, a tool that may serve in restricting outcomes currently skewed.

List the hyperparameter tuned and the values tested and also the rationale why you choose them.

Lasso and Ridge (hyperparameters of key interest)

Alpha: as it is a regularisation strength hyperparameter that controls shrinkage of the coefficients towards zero. A high alpha value leads to greater shrinkage, simplifying models to smaller coefficients.

Normalisation: As it is a boolean value that indicates whether to normalize the features before fitting the model.

KNN (hyperparameter of key interest)

k “n_neighbors”: as it scales the number of nearest neighbours before making its prediction, a feature useful for to interpret some features with reduced correlative data points.

Random Forest (hyperparameters of interest)

“n_estimators”: to determine the number of trees

“max_depth”: to determine the depth of each tree

“min_samples_split”: Minimum number of samples required to split a node

“random_state”: further aided in reproducibility through randomisation

- *Most hyperparameters tuned will be tweaked and experimented with, to place stronger constraints and boundaries to improve accuracy evaluations, though only some will remain in the final experiment notebook.*

List also the models you decided to not train and the reasoning behind it.

Polynomial Regression:

Through the use of a polynomial function. While useful when relationships between the dependent and independent variables is nonlinear, it can be prone to overfitting, especially with higher-order polynomials.

Support Vector Regression (SVR):

Through the use of support vector machines, it can be useful when the relationship between the dependent and independent variables are nonlinear and there are outliers. It was not investigated due to further learning required on the hyper parameter tuning component. For this reason, it may be useful to explore in future model iterations and explorations.

Highlight any model or hyperparameter that may potentially be important for future experiments.

“random_state”:

As this hyperparameter identifies the reproducibility of the information through randomisation. This hyperparameter is most applicable for future iterations, to determine accuracy and authenticity of outputs, which is crucial should deployment be considered.

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

Multivariate Linear Regression with feature engineering:

Model Evaluation:

Mean Squared Error: 378.3021422884047

Mean Absolute Error: 13.670669129146667

Outcome: Does not meet business objective.

Rational for performance:

While an improvement on previous model iterations, the added feature engineering still produces a relatively poor accuracy. This is likely because the data is simply not sufficient, too spread, and too inconsistent with its locations. For example, there is a single incidence/deathrate relationship that far exceeds the critical mass. An example of this may have been the “Geography” feature. Likewise, the features could always have been engineered further and more intricately, accounting for further correlations missed during the initial engineering phase and subsequent assessment.

3.a.
Technical
Performance

Lasso Regression:

Model Evaluation:

Mean Squared Error: 418.81424910000175

Mean Absolute Error: 14.691128601393535

Mean Squared Error (alpha): 378.4500813367334

Mean Absolute Error (alpha): 13.673908566222082

Outcome: Does not meet business objective.

Rational for performance:

As the Lasso regression process shrinks coefficients of lower importance to generate a simpler model (a model that still represents a very complex field of inquiry, cancer mortality rate), the lasso regression may have reduced the data frames too excessively, resulting in less accurate overall interpretation of datapoints. Likewise, it may have capitalised on data features that were larger due to the sporadic spread of features and counties, resulting in higher collection of data points from one region and lowering the weight of data points from others, completely changing the way the relationships interact.

Further experimentation with hyperparameter tuning may serve to improve this accuracy, though after examining the effects with “alpha”, “max_iter” and “normalisation”, these follow through with marginal influence.

Ridge Regression:

Model Evaluation:

Mean Squared Error: 378.78817284813806

Mean Absolute Error: 13.678940554458162

Outcome: Does not meet business objective.

Rational for performance:

The Ridge regression was incorporated to identify multicollinearity through an L2 regression, accounting for large variances between some values. The trouble is, the results predicted are typically not proximal to the actual values, and therefore aren't typically an adequate representation of the real scenario. As some co-linearity seems to be present in particularly similar values and columns, namely the percentages of county residents with specified health coverages, the rational was that the inclusion may help. While among one of the more accurate evaluations, it's perceived the sampling method and subsequent nature of the dataset limits optimisation for this model.

KNN Regression:

Model Evaluation:

Mean Squared Error: 615.7960252587992

Mean Absolute Error: 18.55223602484472

Mean Squared Error (k): 241.01223602348447

Mean Absolute Error (k): 5.364182194616976

Outcome: Does not meet business objective.

Rational for performance:

As KNN approximates independent variable associations through averages of proximal neighbours, the perception to experiment with it seemed correct. This however resulted in a less accurate model than previous iterations. By tuning the hyperparameter “n_neighbors” (a procedure still left in the code), the result instead improved to 241. As some data points exist in small clusters depending on county of origin, the tuning process may have elucidated this feature. The final business objective, however, is yet to be reached.

Random Forest Regression:

Model Evaluation:

Mean Squared Error: 138.86742156107675

Mean Absolute Error: 7.203492753623192

Rational for performance:

As Random Forest estimates by measuring subsamples of data averages, inclusion was necessary to account for the arbitrary nature of the data sampling. This proved to be a strong choice, as the final MSE evaluation yielded the most accurate of all previously assessed models. This was evaluated with hyperparameter random_state, a feature that saw consistency with the final evaluation, though with marginal increases (thus was removed from the final code). Despite performing well, this model still does not meet the business goal, and therefore is interpreted as a need to retrieve more data from the field for further optimisation.

Interpret the results of the experiments related to the business objective set earlier.

3.b.
Business
Impact

As the objective of the project is to develop a predictive model that can **accurately** predict cancer mortality in different US counties based on a 33-feature dataset, given the high,

inaccurate MSE variances following evaluations, it does not seem possible to accurately predict cancer mortality, based on the provided data frame. The most accurate metric delivered was an MSE of 138.9, which still leaves a large room for error. This is especially problematic, as the errors of concern deal with predictions over human lives. The results therefore should not proceed to deployment.

Estimate the impacts of the incorrect results for the business, (some results may have more impact compared to others).

Unfortunately the inability to deliver an accurate model means cancer mortality remains uncategorisable and predictable, meaning the life-conserving benefits mentioned above cannot be delivered. If the Random Forest model were deployed, it would provide inaccurate predictions to patients and families. In a profession where accuracy is paramount in determining likelihood of deathrate (medicine and healthcare), the model may enact unnecessary supply, financial and healthcare outcomes.

List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them.

Hyperparameter tuning – instantiating some parameters were more difficult than others (e.g. `random_forest`) due to their integration method (personal), and longer cpu processing times. Often the problem could be remedied additional research, other times, trial and error.

Coding reliability – human error, resolved with practice and exploring alternate libraries.

Data limitations – more features, would help provide more opportunities to discern correlations and relationships. While not solvable (despite importing external data sources and experimenting with each), it contributed to the final outcome of not meeting the business goal.

Highlight also the issues that may have to be dealt with in future experiments.

Data limitations – more features lead to less overfitting, underfitting and bias. Data limitations can be detrimental such an experiment as they can negatively impact the accuracy and reliability of the results. The issue can only be dealt with sampling more counties, and participants of the counties, comprehensively.

3.c.
Encountered
Issues

4. FUTURE EXPERIMENT

Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.

Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.

4.a.
Key Learning

After conducting the experiment, it became clear that the regression models exercised were not able to deliver the final business goal of an accurate predictor for cancer mortality. Both Lasso and Ridge models were unsuccessful, as was KNN regression, despite hyperparameter tuning. Despite the Random Forest model performing the most accurately, it still was unable to meet the business goal.

A key insight interpreted from the experiment is the necessity for more comprehensive data inclusions, including intricate features such as diet, CO2 emissions, habits, socioeconomic backgrounds (living conditions), and previous medical histories. While participants may be less inclined to part with personal information, posing privacy concerns, results will skew towards biased outcomes, unrepresentative of entire populations. However, even with this additional data, it still may not be possible to achieve an accurate prediction, as cancer mortality is a complex issue that depends on many lifestyle and environmental circumstances, including genetics.

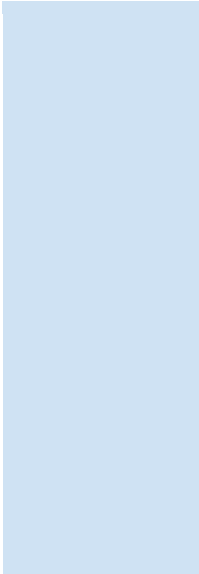
Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly.

4.b.
Suggestions /
Recommendations

The low accuracy model evaluations suggest more experimentation may be worth pursuing through more complex modelling, though is advised to be conducted following the acquisition of a more comprehensive dataset (such as diet, CO2 emissions, habits, socioeconomic backgrounds (living conditions), and previous medical histories).

This would result in no immediate deployment of the model, though the subsequent steps would present an opportunity for revision of existing sampling methods for ones that are uniform, complete (no NaN values) and detailed. The categorical features in some language descriptors could be broken down through the .nlTK natural processing library, to provide more insights about behaviour, choices and environmental determinants of health and health outcomes.

Finally utilising this process as a learning tool to optimise future methods would prove valuable. This would entail an extensive preparation phase prior to investigation. As a data scientist's time is the most valuable faculty in meeting an objective, discerning probable outcomes in meeting the



business goal ahead of schedule would serve to save the business' time and resources.

If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.

<The experiment did not achieve the business objective in accurately predicting cancer mortality rates in US counties, and therefore, it is advised the model is not deployed.

The exploration of data may be repurposed to determine other prediction factors about the population, such as insurance, education, and sales. The outcomes may further be used as a learning tool for constructing future models.>