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 predictive 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 their family members' survivability following diagnosis. Constructing a successful model ultimately reduces the prevalence of cancer or increases patient survival 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 are 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.

1.b. Hypothesis

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

Part A hypothesis:

Certain demographic and socioeconomic factors, such as elevated incidence rates and poverty percentages, are associated with higher cancer mortality rates in US counties; while lower instances of poverty and increased education are associated with lower cancer mortality rates.

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

The hypothesis warrants exploration, as it suggests there may be causality between certain demographic and socioeconomic factors and cancer mortality rates in US counties. Should these relationships be confirmed, it aids 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.c. Experiment Objective

Detail what will be the expected outcome of the experiment.

The expected outcome for **Part A**, is exploration into two features using a Univariate Linear Regression Model for each feature: “incidence rate” and “poverty percent”. As incidence rate is inexplicably tied to cancer manifestation, understanding the relationship of survivability would be deemed critical. It is expected to have a high correlation with the target variable, “deathrate”. Likewise, as “poverty percentage” is a socioeconomic factor that serves to help mitigate a cancerous growth phase through access to treatment. As those with less capital have fewer options, it is expected to be also highly correlated.

If possible, estimate the goal you are expecting.

The anticipated goal would be the construction of a univariate linear regression model which demonstrates a high correlation

between the target feature, “deathrate” and two separate features “incidence rate” and “poverty percentage.”

List the possible scenarios resulting from this experiment.

A significant and highly probable outcome is a lack of noticeable correlation between variables presumably correlated. Should this be the case, resolving the scenario would be to explore other features, or perform feature engineering to discern relationships, if present.

Code miscalculation from human error during the preparation phase is a possible throughout construction. This has the capacity to lead to true positive, or false negative assertions, or presumed statistically significant outcomes and create devastating, real life consequences if deployed. For this reason, a careful eye needs to be had, with persistent, meticulous revision to ensure variables and values aren’t misinterpreted.

Another possible scenario, and the desirable one, is the successful creation and deployment of an accurate model. As human behaviour and is chaotic, variables will never truly be 100% correlated, thus success would be deemed 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.

2.a. Data Preparation

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

Data preparation occupied around 70% of the work invested for Part A. This was paramount, as the steps exercised and interpretations made have the capacity to shape perceptions and decisions for each part.

While the data provided was split already, to make the coding more efficient, the data was first concatenated into a single data frame to be processed altogether. After exploring the dataset, non-integer variables were identified, and in the case of “binnedInc”, was dropped, while “geography” was sorted by state and stored for one-hot-encoding for Part C feature engineering. Missing values were visualised using the “missingno” library, leading to the “PctPrivateCoverageAlone” and “PctEmployed16_Over” missing values being filled with their respective mean at first. After seeing it’s influence on the overall distribution, however, these values

were kept, but were noted as not accurate, as it influenced their overall distributions. “PctSomeCol18_24” however, had too many missing variables to retain in the dataframes.

Duplicate rows were screened for, but none were detected. While the dataset possessed many outliers across all fields, these represented real-world features and must be left in; however, the median age category had to remove any >300, as they weren’t realistic. I proceeded to gauge the impact of leaving the >1000 “incidence rate” outlier. While it was distinct, it remained in the dataset. Features of the dataset were calculated for correlation with each other, to identify any relationships that weren’t strictly tied to the target variable, as these may serve supplementary feature engineer phase in Part C. The strongest positively and negatively correlative features were graphed with the target variable, prior to decided which would suit a Univariate Linear Regression model.

Explain the rationale why you had to perform these steps.

Data preparation is critical in delivering an accurate model. Unnecessary categorical columns were dropped, as they impeded utility in modelling. Missing number columns were dropped, or if marginal, were retained and granted the mean (if normally distributed), purely for additional, rough features to incorporate in Part C, should previous models not meet expectations. Outliers were left in, as they represented real scenarios, minus ages >300. Correlations were constructed prior to training a Univariate Linear Model, to provide the reveal stronger relationships and conserve time.

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

- Removing some noticeable outliers, as they represented real-world scenarios, minus ages exceeding 300.
- Removing the inserted averages for “PctPrivateCoverageAlone” and “PctEmployed16_Over”. While not an accurate representation of the dataset, these features were too useful to drop.
- Inserting more features from other dataframes; as it would have different contextual circumstances with its collection and may shift the results to one that doesn’t reflect real scenarios.
- Removing geography; as it was convertible with one-hot-encoding during Part C, for feature engineering. It remained a non-numerical feature.

Highlight any step that may potentially be important for future experiments

“Taking the top 6 features with the highest correlation with the target variable, and visualising the their relationships.”

By identifying the strongest relationships immediately, it served to distinguish features worthy of modelling. As time is the most valuable asset, prioritising it on the data that mattered, was paramount.

2.b. Feature Engineering

Describe the steps taken for generating features (if any).

While “geography” was reduced to its corresponding States for one-hot-encoding in the data preparation phase, all features were engineered in Part C.

Coverage on these topics are in the corresponding report.

Explain the rationale why you had to perform these steps.

<No features generated for Part A>

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

”binnedInc” was removed, as it was non-numerical and encapsulated another feature already available in the data, “Median Income”.

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

”PctPublicCoverageAlone” – it represents the highest correlation with the target variable and relates to other features – such as “income” and “poverty”.

2.c. Modelling

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

Univariate Linear Regression – a simple, yet strong foundation for visualising relationships between variables. As a starting point, it is helpful for easing into other models.

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

< Linear Regression does not possess hyperparameters to tune.>

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

While six features were singled out as having the strongest correlation. The business objective prioritised performing Univariate Linear Regressions on two (incidence rate and poverty percentage). As incidence rate is inexplicably tied to cancer manifestation, understanding the relationship of survivability would be deemed critical. It is expected to have a high correlation with the target variable, “deathrate”. Likewise, as “poverty percentage” is a socioeconomic factor that serves to help mitigate a cancerous growth phase through access to treatment. As those with less capital have fewer options, it is expected to be also highly correlated.

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

< As a multivariate linear regression will be contain parameter tuning in Part 3, the alpha parameter as a part of the Lasso Regression model. Using this feature will help improve the model’s strength of regularisation by balancing bias and variance. >

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.

3.a. Technical Performance

Score of the relevant performance metric(s).

incidenceRate vs TARGET_deathRate

Training MSE: 599.5
Testing MSE: 663.8

povertyPercent vs TARGET_deathRate

Training MSE: 590.5
Testing MSE: 697.8

Provide analysis on the main underperforming cases/observations and potential root causes.

Data was sourced at different times, in different quantities, with more areas surveyed than others.
Human behaviour is difficult to accurately measure, and each individual possesses faculties that may shape their circumstance. This results in a wide spread in the scatter plot, leading to higher MSE scores. More data and trailing different features may help reduce these MSE scores.

3.b. Business Impact

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

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 variance, it does not seem possible with a univariate linear regression model.

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.

3.c. Encountered Issues

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

Inaccurate model – provide more features (i.e. CO2 output, diet, sun exposure, known carcinogens within areas), alongside more people, with more consistent sampling measures.

Coding reliability – human error, resolved with practice

NaN values – the few missing, fit with the mean; does not represent a real survey.

Geography – data is arbitrarily spread, and in different quantities

Data limitations – more features, would help provide more opportunities to discern correlations and relationships.

Outliers – while noticeable outliers were present, they had to be retained to mimic real-world scanarios; the age values >300 were dropped as it was not realistic.

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

Discerning the impact of Public Coverage and Private Coverage – as missing values were provided the mean of the distribution.

Geography – limited to some US states, with varying degrees and quantities

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.

4.a. Key Learning

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.

While unsuccessful, the exercise provides a learning curve in discerning what is achievable with limited features. While trends in capital, education and decisions lead better or worse cancer mortality rates, it can never be simplified to a simple model. The insights however allow further investigation into further regression models, that may integrate more features. The MSE, while high, can be reduced by conceiving a larger snapshot of the environment of US states, through more data features.

As cancer manifestation is the product of genetic and environmental factors, both dynamics are attainable with a diversity of features. The current features are sparse and limit a perspective of a broad, consistently changing dynamic. For this reason, I do not consider that more experimentation with the current features will yield a model that could reduce lifestyle factors to predictability.

4.b. Suggestions / Recommendations

Given the results achieved and the overall objective of the project, list the potential next steps and experiments.

As Univariate Linear regression is a simple model, the logical next step would be to integrate more features through multiple regression. The large MSE achieved in the first step serves as a foundation for comparison of subsequent steps.

The next steps would be to prepare the data for Multivariate regression, and then subsequently carry out the model.

For each of them assess the expected uplift or gains and rank them accordingly.

Multivariate Linear Regression – wider snapshot of a complex, multifaceted set of interacting data.

Feature Engineering – more features to add to existing data, which may reveal further insights and correlative features

Further regression models, e.g. lasso and random forest – alternatives, that operate different paths at predicting datasets.

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 required outcome, therefore it is advised the model is not deployed, rather used as a means to learn from, so that future models may not encounter similar issues with crafting a "reductionist narrative" of lifestyle factors.>