UNIVERSITY OF TECHNOLOGY, SYDNEY

Machine Learning - Cancer Mortality Prediction

Machine Learning Data-Driven Insights: Predicting Cancer Mortality Rates Using Census Data in US Counties

Monali Patil

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) commonly known as CRISP-DM, is a commonly utilized methodology for data science projects (HOTZ, 2023).

The subsequent activities are performed as part of the learning for this assignment project.

1. Business Understanding

This project aims to develop a regression model that predicts cancer mortality rates in counties across the United States using census data. The given dataset is unified by aggregating census data from across the United States. The primary objective is to explore the relationships between various demographic and medical factors and their impact on cancer mortality rates. This analysis can provide valuable insights for public health planning and resource allocation.

The objective of this project is to build a regression model that accurately predicts cancer mortality rates based on various features extracted from the US census data. By performing data exploration, cleaning, and feature selection, identify the most relevant factors influencing mortality rates and develop a model that can be used to predict these rates on unseen data. This model could help in understanding key determinants of cancer mortality and aid in targeted public health interventions.

Additionally, the resulting model can also be used to obtain valuable insights into the associations between various factors and variables that impact cancer mortality rates. This information can then be employed by healthcare professionals (both public and private), policymakers, and researchers to create better strategies for decreasing cancer mortality rates. Additionally, informed decision-making processes can be implemented to enhance health results and services.

Business Problem:

The primary business problem is to build a (Supervised Machine Learning) Univariate/Multivariate Linear Regression model that can accurately predict cancer mortality rates using data related to counties in the United States. The model aims to identify the most significant individual or multiple features that affect cancer mortality, providing insights that could be used for public health decision-making and resource allocation.

2. Data Understanding

1] Describe the Data

The dataset provided is consolidated from US census data and consists of 33 features related to demography and medical information at the county level such as medical coverages of people, their average age, education, employment. The dataset is split into two parts:

- Training Set: cancer_us_county-training.csv - Used to train the regression model.
- Testing Set: cancer_us_county-testing.csv - Used to evaluate the performance of the model.

Data Dictionary

- TARGET_deathRate: Mean per capita (per 100,000) cancer mortalities.

- avgAnnCount: Mean number of reported cases of cancer diagnosed annually.
- avgDeathsPerYear: Mean number of reported mortalities due to cancer.
- incidenceRate: Mean per capita (per 100,000) cancer diagnoses.
- medianIncome: Median income per county.
- popEst2015: Population estimate of the county in 2015.
- povertyPercent: Percent of the population in poverty.
- studyPerCap: Per capita number of cancer-related clinical trials per county.
- binnedInc: Median income per capita binned by decile.
- MedianAge: Median age of county residents.
- MedianAgeMale: Median age of male county residents.
- MedianAgeFemale: Median age of female county residents.
- Geography: County name.
- AvgHouseholdSize: Mean household size of the county.
- PercentMarried: Percent of county residents who are married.
- ...and other features related to education, employment, insurance coverage, race, and birth rates.

In the dataset, most attributes are of an integer(int64), or decimal(float) data type and only two are of string/character(object) data type. The following table provides a short description of the datasets, their datatypes and attribute names and descriptions.

Training Set

Sr. No.	Attibute Name	IsNull	Datatype	Description
1	TARGET_deathRate	0	float64	Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)
2	avgAnnCount	0	float64	Mean number of reported cases of cancer diagnosed annually(*a*)
3	avgDeathsPerYear	0	int64	Mean number of reported mortalities due to cancer(*a*)
4	incidenceRate	0	float64	Mean *per capita* (100,000) cancer diagoses(*a*)
5	medianIncome	0	int64	Median income per county (*b*)
6	popEst2015	0	int64	Population of county (*b*)
7	povertyPercent	0	float64	Percent of populace in poverty (*b*)
8	studyPerCap	0	float64	*Per capita* number of cancer-related clinical trials per county (*a*)
9	binnedInc	0	object	Median income per capita binned by decile (*b*)
10	MedianAge	0	float64	Median age of county residents (*b*)
11	MedianAgeMale	0	float64	Median age of male county residents (*b*)
12	MedianAgeFemale	0	float64	Median age of female county residents (*b*
13	Geography	0	object	County name (*b*)
14	AvgHouseholdSize	0	float64	Mean household size of county (*b*)
15	PercentMarried	0	float64	Percent of county residents who are married (b)
16	PctNoHS18_24	0	float64	Percent of county residents ages 18-24 highest education attained: less than high school (*b*)
17	PctHS18_24	0	float64	Percent of county residents ages 18-24 highest education attained: high school diploma (*b*)
18	PctSomeCol18_24	1826	float64	Percent of county residents ages 18-24 highest education attained: some college (*b*)
19	PctBachDeg18_24	0	float64	Percent of county residents ages 18-24 highest education attained: bachelor's degree (*b*)
20	PctHS25_Over	0	float64	Percent of county residents ages 25 and over highest education attained: high school diploma (*b*)
21	PctBachDeg25_Over	0	float64	Percent of county residents ages 25 and over highest education attained: bachelor's degree (*b*
22	PctEmployed16_Over	122	float64	Percent of county residents ages 16 and over employed (*b*)
23	PctUnemployed16_Over	0	float64	Percent of county residents ages 16 and over unemployed (*b*)
24	PctPrivateCoverage	0	float64	Percent of county residents with private health coverage (*b*)
25	PctPrivateCoverageAlone	500	float64	Percent of county residents with private health coverage alone (no public assistance) (*b*)
26	PctEmpPrivCoverage	0	float64	Percent of county residents with employee-provided private health coverage (*b*)
27	PctPublicCoverage	0	float64	Percent of county residents with government-provided health coverage (*b*)
28	PctPubliceCoverageAlone	0	float64	Percent of county residents with government-provided health coverage alone (*b*)
29	PctWhite	0	float64	Percent of county residents who identify as White (*b*)
30	PctAsian	0	float64	Percent of county residents who identify as Asian (*b*)
31	PctOtherRace	0	float64	Percent of county residents who identify in a category which is not White, Black, or Asian (*b*)
32	PctMarriedHouseholds	0	float64	Percent of married households (*b*)
33	BirthRate	0	float64	Number of live births relative to number of women in county (*b*)
34	ID	0	int64	

Table 1: Training dataset highlighting null values.

Testing Set

Sr. No.	Attibute Name	IsNull	Datatype	Description
1	TARGET_deathRate	0	float64	Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)
2	avgAnnCount	0	float64	Mean number of reported cases of cancer diagnosed annually(*a*)
3	avgDeathsPerYear	0	int64	Mean number of reported mortalities due to cancer(*a*)
4	incidenceRate	0	float64	Mean *per capita* (100,000) cancer diagoses(*a*)
5	medianIncome	0	int64	Median income per county (*b*)
6	popEst2015	0	int64	Population of county (*b*)
7	povertyPercent	0	float64	Percent of populace in poverty (*b*)
8	studyPerCap	0	float64	*Per capita* number of cancer-related clinical trials per county (*a*)
9	binnedInc	0	object	Median income per capita binned by decile (*b*)
10	MedianAge	0	float64	Median age of county residents (*b*)
11	MedianAgeMale	0	float64	Median age of male county residents (*b*)
12	MedianAgeFemale	0	float64	Median age of female county residents (*b*
13	Geography	0	object	County name (*b*)
14	AvgHouseholdSize	0	float64	Mean household size of county (*b*)
15	PercentMarried	0	float64	Percent of county residents who are married (b)
16	PctNoHS18_24	0	float64	Percent of county residents ages 18-24 highest education attained: less than high school (*b*)
17	PctHS18_24	0	float64	Percent of county residents ages 18-24 highest education attained: high school diploma (*b*)
18	PctSomeCol18_24	0	float64	Percent of county residents ages 18-24 highest education attained: some college (*b*)
19	PctBachDeg18_24	0	float64	Percent of county residents ages 18-24 highest education attained: bachelor's degree (*b*)
20	PctHS25_Over	0	float64	Percent of county residents ages 25 and over highest education attained: high school diploma (*b*)
21	PctBachDeg25_Over	0	float64	Percent of county residents ages 25 and over highest education attained: bachelor's degree (*b*
22	PctEmployed16_Over	30	float64	Percent of county residents ages 16 and over employed (*b*)
23	PctUnemployed16_Over	0	float64	Percent of county residents ages 16 and over unemployed (*b*)
24	PctPrivateCoverage	0	float64	Percent of county residents with private health coverage (*b*)
25	PctPrivateCoverageAlone	109	float64	Percent of county residents with private health coverage alone (no public assistance) (*b*)
26	PctEmpPrivCoverage	0	float64	Percent of county residents with employee-provided private health coverage (*b*)
27	PctPublicCoverage	0	float64	Percent of county residents with government-provided health coverage (*b*)
28	PctPubliceCoverageAlone	0	float64	Percent of county residents with government-provided health coverage alone (*b*)
29	PctWhite	0	float64	Percent of county residents who identify as White (*b*)
30	PctAsian	0	float64	Percent of county residents who identify as Asian (*b*)
31	PctOtherRace	0	float64	Percent of county residents who identify in a category which is not White, Black, or Asian (*b*)
32	PctMarriedHouseholds	0	float64	Percent of married households (*b*)
33	BirthRate	0	float64	Number of live births relative to number of women in county (*b*)
34	ID	0	int64	

Table 2: Testing dataset highlighting null values.

2] Explore Data

The data was explored utilizing the below functionalities of the pandas and seaborn data visualisation library.

Exploratory Data Analysis

df.head()	For checking some datapoints of the dataset.
df.sample()	For checking sample/any 5 rows of the dataset.
df.shape	For describing the dimension/number of rows and columns of the dataset.
df.columns	For checking attributes names of the dataset.
df.info()	For checking attributes summary information(features datatypes) of the dataset.
df.describe()	For checking the summary statistics the dataset.
df.describe(include='all')	For describing summary statistics for all datatype variables of the dataset.
df.isnull().sum()	To identify if any null values in the dataset.

```
* To drop the null values used drop() function.
df_train.drop(['PctSomeCol18_24'], axis=1, inplace=True)

* To fill the null values with mean value.
df_train['PctEmployed16_Over'].fillna(df_train['PctEmployed16_Over'].mean(), inplace=True)

* To check if any missing values, generate the heaatmap with below syntax.
sns.heatmap(df_train.isnull(), yticklabels=False, cbar=True)

* To find the relationship between the target and other different attributes, plotted Heatmap for Correlation.
sns.heatmap(df_train.corr(), annot=True, cmap='summer_r')

* To check the distribution of the datapoints of features below histograms are used.
df_continuous_features.hist(figsize=(10,10))

* For examining outliers present in the continuous features below boxplots are used.
sns.boxplot(x='incidenceRate', data=df_continuous_features, ax=axes[0], color='orange')

* To view the data distribution of the any continuous features with the target variable pairplots are used.
sns.pairplot(df_train, x_vars=['incidenceRate'], y_vars='TARGET_deathRate', height=5)
```

Once the initial data exploration stage is completed, the data is prepared and processed to make it suitable for use with algorithms. As it's aggregated data it appears that individual identities were cleaned, and the quality of data was acceptable to be used as no issues were encountered except null values.

3. Data Preparation

One of the important factors while working on the supervised machine learning models is that the machine learning algorithms are not capable of dealing with missing values in continuous attributes or any other attributes. Therefore, it is necessary to either address the issue of missing values or eliminate the corresponding columns altogether. And the usual way is to substitute it with mean value as it is an average of the ranges of values present in the attribute.

However, from the 'PctSomeCol18_24' feature more than 70% of the records are missing (only 612 recorded out of 2438), so it is impractical to fill them with mean values. Therefore, removing this feature from both datasets. And the rest features' null values were replaced with their mean.

For Part C, feature engineering is performed, and additional information is extracted from 'binnedInc' attribute by performing below steps and stored in 'binnedInc1' and 'binnedInc2'.

```
# Removing beginning and tailing brackets '()' and '[]' from the 'binnedInc' feature from training set.
df_train['binnedInc'] = df_train['binnedInc'].str.strip('[]').astype(str)
df_train['binnedInc'] = df_train['binnedInc'].str.strip('()').astype(str)

# Splitting the two set of information into two new columns 'binnedInc1' and 'binnedInc2' from training set.
df_train[['binnedInc1','binnedInc2']] = df_train['binnedInc'].str.split(',',expand=True)

# Removing beginning and tailing brackets '()' and '[]' from the 'binnedInc' feature from testing set.
df_test['binnedInc'] = df_test['binnedInc'].str.strip('[]').astype(str)
df_test['binnedInc'] = df_test['binnedInc'].str.strip('()').astype(str)

# Splitting the two set of information into two new columns 'binnedInc1' and 'binnedInc2' from testing set.
df_test[['binnedInc1','binnedInc2']] = df_test['binnedInc'].str.split(',',expand=True)
```

4. Modelling

Below models are built, trained and tested as part of this learning process.

Sr. No.	Assignment	Model Employed	Algorithm Utilised	Independent Features
1	Part A - I	Univariate Linear Regresson	LinearRegression	incidenceRate'
2	Part A - II	Univariate Linear Regresson	LinearRegression	PctPublicCoverageAlone'
3	Part B - Exp1	Multivariate Linear Regression	LinearRegression	incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage'
4	Part B - Exp2	Multivariate Linear Regression	LinearRegression	All continuous features except 'ID', 'binnedInc' and 'Geography' as these are object/string/categorical features.
5	Part C - Exp1	Multivariate Linear Regression	Lasso Linear Regression	incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2'
6	Part C - Exp2	Multivariate Linear Regression	Ridge Linear Regression	incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2'
7	Part C - Exp3	Multivariate Linear Regression	Elasticnet Linear Regression	incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2'
8	Part C - Exp4	Multivariate Linear Regression	KNN with 50 neighbours and Euclidean Distance	incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2'

Modules for all the algorithms are imported from sklearn library.

Although Multivariate Linear Regression is used for assignment learning purposes, it is an appropriate choice for both univariate and multivariate linear regression because the target variable, "TARGET_deathRate," and all the chosen single/multiple independent variables are continuous and can potentially have infinite values. Therefore, Multivariate Linear Regression is a suitable model choice for all the experiments.

Some of the Key Decisions Made about the Modeling

1. For Part A both the features were selected based on signifiacnt high correlation with target variable derived from the correlation heatmap chart, and datapoints distribution against target variable and identiify if there is any linear pattern.

2. For Part B, experiment 1, the 6 predictors were choosen based on same, strong correlation with target variable derived from the correlation heatmap graph and by checking if relatively linear distribution.

3. For Part C, used the learning from part B and further tried to reduce overfitting using different algorithms to add penalty and to reduce weight by any specific information thereby increasing the performance by generalising the model.

4. Couldn't employ hyperparameter due to time constraints and used models inbuilt parameters to regularise the model but would work on this in future experiments.

5. Model Evaluation

To evaluate the algorithm's performance utilised below performance metircs.

1. Baselines Performance.
2. Mean Square Error (MSE)
3. Mean Absolute Error (MAS)
4. Root Mean Square (RMSE)

Baseline performance acts as the simplest model that always predicts the same baseline value which helps to evaluate the performance of the trained linear regression models and ensure that we are making progress in the right direction.

The MSE calculates the average of the squared differences between the predicted and actual values. Since the unit of values is doubled, the error is emphasised. Therefore, to mitigate this effect, the RMSE is used for evaluation below, which cancels out the squaring effect and brings the unit of measurement back to its original scale.

Below are the Performance Metrics of the trained linear regression algorithms.

- **Experiment: Part A - I (Univariate Linear Regression)**

Performance Metrics of Univariate Linear Regression with 'incidence Rate' feature.

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	612.89	19.3498	24.7566
Validation	614.1857	19.2027	24.78277
Testing	622.488	19.7412	24.9497

- **Experiment: Part A - II (Univariate Linear Regression)**

Performance Metrics of Univariate Linear Regression with 'PctPublicCoverageAlone' feature.

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	579.9413	18.3111	24.0819
Validation	719.2265	19.9816	26.8184
Testing	698.2437	19.2253	26.4243

- As the difference between the RMSE scores for 3 sets of experiment 1 is relatively small, it suggests that the model is performing consistently across the three sets and is not overfitting to the training data and this univariate linear regression model with a single 'incidenceRate' independent feature is generalizing well to new/unseen data.

- The difference between the RMSE scores for 3 sets of experiment 2 is small and it appears that the model's performance is decent but not ideal. Therefore, it appears relying solely on the 'PctPublicCoverageAlone' independent feature is inadequate to make precise predictions about cancer mortality rates.

- The model from the first experiment performed consistently for all the 3 sets and is generalised fairly to predict cancer mortality rate, whereas the model from the second experiment is sightly overfitting the training set.

- **Part B: Experiment 1 (Multivariate Linear Regression)**

Performance Metrics of Multivariate Linear Regression (with chosen 6 features)

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	397.0357	14.8642	19.9257
Validation	367.6438	14.672	19.174
Testing	453.8926	15.758	21.3047

- **Part B: Experiment 2 (Multivariate Linear Regression)**

Performance Metrics of Multivariate Linear Regression (with all features except a few with object)

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	343.6132	13.922	18.5368
Validation	482.1358	14.9528	21.9575
Testing	417.823	15.0174	20.4407

- The difference between the RMSE scores of both experiments for all three sets is relatively small, and it suggests that the model is slightly overfitting to some extent on the training data as there be some datapoints that are specific, and the model couldn't be enough generalised to predict unseen observation.

- Moreover, the model from experiment 1 is performing slightly better than experiment 2, on the testing data compared to the training data. Therefore, it may appear that selecting multiple independent features based on correlation values shows relatively better performance suggesting a reasonable and informed selection of predictors than training a model with all continuous variables.

- However, the models are not sufficiently generalized, and there are still features with observations that are specific, so it is necessary to adjust the model hyperparameters,

regularizing the model to reduce its specification and prevent overfitting based on business goals.

- Part C: Experiment 1 (Multivariate Lasso Linear Regression)

Performance Metrics of Lasso Linear Regression.

Lasso model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	392.520168	14.832555	19.812122
1	Validation	370.662380	14.663188	19.252594
2	Testing	442.152615	15.582073	21.027425

- Part C: Experiment 2 (Multivariate Ridge Linear Regression)

Performance Metrics of Ridge Linear Regression.

Ridge model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	379.603859	14.746589	19.483425
1	Validation	527.413031	15.768654	22.965475
2	Testing	454.341259	15.772174	21.315282

- Part C: Experiment 3 (Multivariate Elasticnet Linear Regression)

Performance Metrics of Elasticnet Linear Regression.

Elasticnet model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	397.359370	14.891666	19.933875
1	Validation	370.779585	14.737633	19.255638
2	Testing	452.724169	15.781192	21.277316

- Part C: Experiment 4 (Multivariate KNN Linear Regression)

Performance Metrics of KNN with 50 neighbours and Euclidean Distance.

KNN model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	406.660094	15.342533	20.165815
1	Validation	430.380194	15.683000	20.745607
2	Testing	508.812250	16.657327	22.556867

  - The difference between the RMSE scores in comparison with multivariate linear regression from part B informs that experiment 1 with the Lasso algorithm and experiment 3 with the ElasticNet has slightly improved performance and are regularised to a small margin, while the

other two experiments using Ridge and KNN with 50 neighbours and Euclidean Distance have only shown a negligible impact. And further testing with regularisation is essential to reduce the errors.

- The line charts from these experiments inform that some data points are distant from the line, which indicates a need for further analysis and validation of those observations. A different approach, such as manual intervention, may be necessary when predicting those data points.

- Therefore, to enhance the model's ability to generalize, it is significant to further consult businesses and understand those observations, their accuracy and business perspectives and benchmarks.

• Deployment

To deploy a linear regression algorithm successfully in the operational environment, it is recommended to follow the below steps.

1) Scale and transform the model to handle big datasets.
2) Select a suitable deployment environment cloud or on-premises.
3) Transform the model for production settings ensuring compliance with various security, ethical and privacy and obligator guidelines.
4) Conduct testing.
5) Monitoring and updating the model periodically with new data.
6) Provide documentation for usage.
7) Review the model's performance and retrain it accordingly.

It is crucial to examine ethical, privacy, and security guidelines during every stage of the process in order to guarantee that no one is harmed. For instance, applying linear regression models from the assignment learning might require and involve handling sensitive data like medical and financial information. So, it is essential to implement data security and privacy measures to prevent unauthorised access or theft.

Some of the other issues are data biases, transparency, and legal and compliance problems. Therefore, it is essential to prioritize ethical and privacy considerations throughout the entire process of deploying a linear regression model in production which involves selecting appropriate data sources, rigorous testing, and validating the model, and regularly monitoring and maintaining it to ensure fairness, accuracy, and trustworthiness.

• Results Achieved

Please refer to the Model Evaluation section for the results obtained.

• Issues Faced

Below are some of the major challenges faced.

1. Learned that the entire dataset was split into 10 bins/sections, but it was unclear if it was based on 'medianIncome' or any other feature.
2. Since the data is aggregated and not actual values, it is challenging to obtain precise insights and understanding.
3. MSE score interpretation but it was clear after visitng lecture notes.

6. Model Summary and Business Suggestions:

Below are some of the recommendation.

1. Would like to perform further testing on the model from Part C with the hyperparameters and evaluate the performance.
2. Also would introduce other functionality from learning point of view by analysing recall, accuracy and try different approaches to examine the model's predictions visually to compare them to the ground truth.

According to the performance metrics, the only model in PartA-I that performed well was in predicting the cancer mortality rate compared to the others. However, since predicting a life-threatening condition such as cancer mortality rate using a single predictor is not acceptable in medical applications, deploying the model in production is not recommended. Additionally, the other models in the experiments are overfitting, so they require further analysis and testing.

- References

HOTZ, N. (2023, January 19). *What is CRISP DM?* Datascience. Retrieved March 31, 2023, from https://www.datascience-pm.com/crisp-dm-2/

So, A. (2023, February 23). *Machine Learning Algorithms and Applications* [Lab]. https://colab.research.google.com/drive/1ZU6hhuvbQ21_BPEGluVAA8p8YhDOPUCf?usp=share_link#scrollTo=OoZFiqIhprGQ

So, A. (2023, March 2). *Machine Learning Algorithms and Applications* [Lab]. https://colab.research.google.com/drive/12lQP6UUXu557PH7rMVs9zCcz5GxRDw-i?usp=share_link#scrollTo=mnaS7_m5reCk

So, A. (2023, March 9). *Machine Learning Algorithms and Applications* [Lab]. https://colab.research.google.com/drive/1QksDWQa8l-prX0lEPF0z7gwfDVTt0NQY?usp=share_link#scrollTo=HpYosTXg2qIz