International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

Analysis and Estimation of Child Mortality and the Influence of Maternal Care on it

Rutuja Mandapmalvi[1], Vishwajeet Ohal[2], Atharva Tonape[3], Ganesh Madhikar[4]

[1-3]Student, Dept. of Computer Engineering, Sinhgad College of Engineering, Maharashtra, India
[4]Assistant Professor, Dept. of Electronics and Telecommunication Engineering, Sinhgad College of Engineering, Maharashtra, India.

Abstract - The project analyzed data of National Family Health Survey (NFHS) 2015-16, conducted as a collaborative project of the International Institute for Population Sciences(IIPS), to unravel the underlying relationships between Infant Mortality Rate (IMR) and the influences of one of the most essential factors, Maternal Care, on it. A comprehensive exploratory analysis was conducted to show significant relationships. The issue of missing values was solved by imputing them with the average of the particular attribute of data. The Linear Regression machine learning model was used to detect and find correlations underlying the data. A comparison between 5 different variants of Linear Regression models was carried out to assess which one performed best given the current scenario. These models were used to make predictions, which can subsequently be used to make changes and reforms to planning in Healthcare Infrastructure Units.

Key Words: **Machine learning, regression, supervised learning, healthcare**

1.INTRODUCTION

Each year in India, there are approximately 28 million pregnancies, 26 million live births, 67 000 maternal deaths and a million neonatal deaths. There are about 42 infant deaths per 1000 live births. Infant mortality is the death of young children under the age of 1. This death toll is measured by the infant mortality rate (IMR), which is the probability of deaths of children under one year of age per 1000 live births. The under-five mortality rate, which is referred to as the child mortality rate, is also an important statistic, considering the infant mortality rate focuses only on children under one year of age.

1.1 Problem Statement

This Project aims to unravel the underlying relationships between Infant Mortality Rate (IMR) and the influences of one of the most essential factors, Maternal Care, on it. Infant Mortality Rate (IMR) gives us the key information about maternal and infant health, the infant mortality rate is an important marker of the overall health of a society.

1.2 Objective

This project aims to analyze data related to Infant Mortality Rate (IMR) and various factors of Prenatal Care from the Ministry of Health and Family Welfare of the Government of India and to build an interactive dashboard giving information about the IMR in India through its various factors. Python was used for data cleaning, data analysis and data visualization. A home page was built showing a map of India. Hovering over states gives key details about the IMR in those states and clicking over it gives more specifics about the IMR scenario in that state. The project also has a prediction mechanism using regression which gives the IMR based on the factors.

1.3 Relevance

This project aims to aid managers of healthcare institutions to make better decisions while understanding the relationships between the different aspects of maternal care that affect infant mortality rate. Using the application, the managers can get precise predictions by entering a set of parameter values to estimate how certain changes in values affect the IMR. Following this, they can make real-life decisions to minimize the IMR within their locality.

2. Methodology

Given that we need to map the relationship between one or more independent features with a dependent feature, it was clear that we needed to use a regression model to establish this. There are a variety of regression models available, with each one having its advantages and limitations. Hence, a need to try and test various models that would work well with the dataset at hand was felt. The dataset we used had mostly numeric data with moderate data density, so we decided to test the following models:

Simple Linear Regression

Multiple Linear Regression

Lasso Regression

Polynomial Regression

Ridge Regression

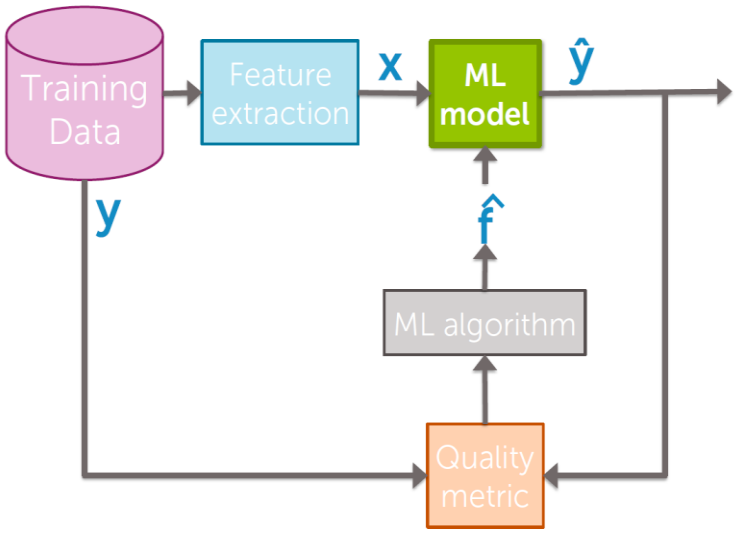

Fig -2.1: Workflow for Machine Learning [2]

2.1 Linear Regression

One is a predictor or independent variable and other is response or dependent variable. It looks for statistical relationships but not deterministic relationships. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other.

Linear regression assumes that a straight line exists in the scatter plot of the independent variable versus the dependent variable. This line is then used to predict the value of the dependent variable given any value of the independent variable. This is accomplished by calculating the y-value corresponding to the given x-value using the general line formula below. Here, w_0 is the weight for the independent variable while b is the y-intercept.

$$y = w_0 x + b$$

2.2 Multiple Regression

Multiple Regression follows a similar approach to Linear Regression but differs in the way that more than one independent variable is used to establish relationships to the dependent variable.

For the project, the following list of independent features was selected:

```
['first_trimester_check',
 'at_least_4_checks',
 'tetanus_vaccination_mothers',
 'folic_acid_consumed',
 'full_care',
 'MCP_card',
 'postnatal_care',
 'financial_assistance',
 'avg_expenditure',
 'home_post_partum_check',
 'check_2_days']
```

Fig -2.2.1: List of features of multiple regression model

After training the model on a 0.8 train-test split, the following coefficients were obtained

Table -2.2.1 Coefficients for Multiple Regression

name	index	value
(intercept)	None	47.291528148195
check_2_days	None	0.3059183358538693
folic_acid_consumed	None	0.22941927016792596
financial_assistance	None	0.1956684443905419
first_trimester_check	None	0.1650082976657481
home_post_partum_check	None	0.0926350065263792
tetanus_vaccination_mothers	None	0.03307424216042351
avg_expenditure	None	-0.0008790121063311567
institutional_births	None	-0.04024881459400361
MCP_card	None	-0.043271422582478
at_least_4_checks	None	-0.1801146493082395
full_care	None	-0.2538911515360698
postnatal_care	None	-0.31123387473209363

Training a Multiple Regression on model gave a model with substantially better accuracy with max error 12.86, RMSE 5.83

2.3 Lasso Regression

Choosing the right features is a crucial task, and can vary the result of predictions by a large degree with a change in selected features. Generally, features are selected using one of two approaches. Greedy approach, where features are cycled through iteratively or Regularization. Lasso Regression of L1 Regularized Regression is an example of the latter approach.

When performing Lasso Regression, the cost function is modified as follows.

This results in coefficients of features dropping to exactly zero, which is equivalent to not considering that feature entirely.

When Lasso regression was performed on this particular dataset, it was seen that no coefficients were set to zero without decreasing the RMSE values. Hence, we proceed by concluding that all the coefficients have a significant impact on the target value of IMR.

2.4 Polynomial Regression

Polynomial Regression uses higher-order values of features in order to capture more depth than what Linear Regression can offer. Using this, quadratic, cubic, and n^{th}-order relationships can be extracted.

```
#Function for getting powers of values of given features for running Polynomial Regression.
#polynomial_sframe accepts dataset (SFrame), list of features (string), and degree of polynomial (integer),
#target column name(string)
# and returns an SFrame
def polynomial_sframe(data, features, degree, target=None):
    # assume that degree >= 1
    poly_sframe = turicreate.SFrame()
    for feature in features:
        # and set poly_sframe['power_1'] equal to the passed feature
        poly_sframe[feature + '_power_1'] = data[feature]
        if degree > 1:
            for power in range(2, degree+1):
                name = feature + '_power_' + str(power)
                poly_sframe[name] = data[feature] ** power
    if target:
        columns = poly_sframe.column_names()
        poly_sframe[target] = data[target]
        return (poly_sframe, columns)
    return poly_sframe
```

Fig -2.1 Algorithm for Computing Polynomial Features [2]

Training the model gave the following coefficients

name	index	value
first_trimester_check_power_1	None	-1.0570164882005484
full_care_power_1	None	-0.5336571943155781
folic_acid_consumed_power_1	None	-0.44630405266685574
home_post_partum_check_power_2	None	-0.19753205019617529
financial_assistance_power_1	None	-0.10272418889199461
check_2_days_power_2	None	-0.014685257170431275
postnatal_care_power_2	None	-0.007741936861624842
tetanus_vaccination_mother...	None	-0.0038896951207166093
MCP_card_power_2	None	-0.0031971121828954364
avg_expenditure_power_1	None	-0.0028745337094044365
institutional_births_power_2	None	-0.0019790097125269033
at_least_4_checks_power_2	None	-0.0015933263646767798
avg_expenditure_power_2	None	2.2768007717221265e-07
full_care_power_2	None	0.002015688182893721
financial_assistance_power_2	None	0.004027614551974532
folic_acid_consumed_power_2	None	0.010169369217329739
first_trimester_check_power_2	None	0.011448713910971336
tetanus_vaccination_mother...	None	0.04640004084616889
at_least_4_checks_power_1	None	0.059062443196514845
institutional_births_power_1	None	0.1900111157995411
MCP_card_power_1	None	0.5136434274481652
postnatal_care_power_1	None	0.6839977511798326
check_2_days_power_1	None	1.2094461640691456
home_post_partum_check_power_1	None	1.3568363455729346
(intercept)	None	48.749352609845424

Table -2.1 Coefficients of Polynomial Regression

Using these coefficients, evaluating the model on test data gave the following results

Max Error 20.81, RMSE 11.40. This accuracy is worse than the multiple regression model and so, we make an attempt to fine tune this model using L2 Regularized Regression as discussed below.

2.5 Ridge Regression

Ridge Regression or L2 Regularized Regression aims to reduce the effect of overfitting by penalizing coefficients with higher magnitude. Overfitting is the problem wherein the trained model makes predictions too close to the data points in the training dataset and does not generalize well to new query points that it may have not observed while training. To remedy this problem, the cost function is modified as below.

$$\text{Total cost} = \underbrace{\text{measure of fit}}_{\text{RSS}(\mathbf{w})} + \lambda \underbrace{\text{measure of magnitude of coefficients}}_{||\mathbf{w}||_2^2}$$

A technique called Cross-Validation was used to generate multiple validation sets in order to find out an appropriate value for tuning parameter lambda. This resulted in the following trend of L2 Lambda vs RMSE and a value of 250 was selected near the elbow of the curve.

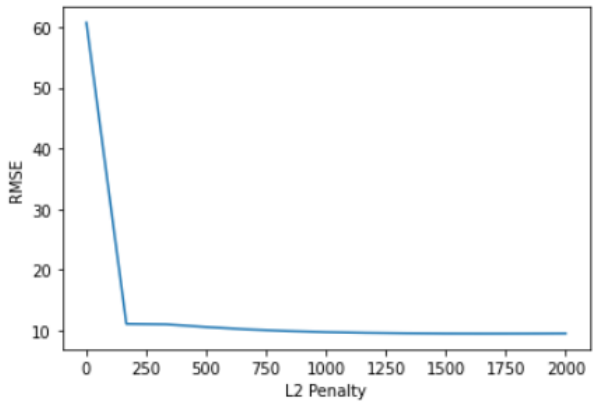

Fig -2.5.1 L2 Lambda vs RMSE for Ridge Regression

Although the RMSE of this model turned out to be better than both the aforementioned Simple Linear Regression and the Polynomial Regression models, with a value of 7.42, it did not outperform the Multiple Regression model. Hence, the Multiple Regression model was chosen as the best model and suitable for Deployment

3. Conclusion

The undertaking analyzed information from National Family Health Survey (NFHS) 2015-16, to solve the essential relationships among Infant Mortality Rate and the impacts of one of the most vital factors, Maternal Care, on it. Exploratory Analysis unearthed some crucial correlations between IMR and Prenatal Care. Compared with no care, prenatal care was associated with lower IMR. These results demonstrate prenatal care is associated with lower IMR. The issue of missing values was solved by replacing them with the average of the particular attribute of data. For this application, the Multiple Regression model provided the most accurate predictions. The Linear Regression machine learning model was preferably used to detect and find correlations underlying the data. A comparison between 5 different variants of Linear Regression models was carried out to assess which one performed best given the current scenario. Other models perform better with datasets of larger size. The limited number of data points led to a few coefficients having improper values. These models had been used to make predictions, which could finally be used to make modifications and reforms to making plans for healthcare infrastructure units and personnel across the nation.

4. References

[1] Dataset obtained from an open government data (OGD)platform of India: https://data.gov.in/resources/all-india-level-and-state-wise-key-indicators-nfhs-3-and-nfhs-4

[2] A machine learning course by coursera to learn various types of machine learning and their applications: https://www.coursera.org/specializations/machine-learning

[3]A youtube channel referred for the statistical study: https://www.youtube.com/user/jbstatistics

[4]Article for reference: obtained from the following website: https://medium.com/topic/machine-learning

[5]https://www.tutorialspoint.com/flask/flask_application.htm

[6] Muthukrishnan, R; Rohini, R (2016). [IEEE 2016 IEEE International Conference on Advances in Computer Applications (ICACA) - Coimbatore, India (2016.10.24-2016.10.24)] 2016 IEEE International Conference on Advances in Computer Applications (ICACA) - LASSO: A feature selection technique in predictive modeling for machine learning. , (), 18–20. doi:10.1109/ICACA.2016.7887916

[7] Library of Congress Cataloging-in-Publication Data Names: Montgomery, Douglas C., author. | Peck, Elizabeth A., 1953- author. | Vining, G. Geoffrey, 1954- author. Title: Introduction to linear regression analysis / Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. Description: Fifth edition. | Hoboken, New Jersey: Wiley. [2020] | Series: Wiley series in probability and statistics Includes bibliographical references and index. Identifiers: LCCN 2020034055 (print) | LCCN 2020034056 (ebook) | ISBN 9781119578727 (hardback) | ISBN 9781119578741 (adobe pdf) | ISBN 9781119578758 (epub) Subjects: LCSH: Regression analysis. Classification: LCC QA278.2 M65 2020 (print) | LCC QA278.2 (ebook) | DDC 519.5/36-dc23 LC record available at https://lccn.loc.gov/2020034055 LC ebook record available at https://lccn.loc.gov/2020034056 Cover Design: Wiley Cover Images: Abstract marbled background, blue marbling wavy lines oxygen/Getty Images Linear Regression analysis graph Courtesy of Douglas C. Montgomery. Set in 10/12pt Times TenRoman by SPi Global, Pondicherry, India