**QMB6304.002S20: - Analytical Methods for Business.**

**A close up of a logo

Description automatically generated**

**Airbnb – Austin Yield Prediction**

**Submitted by: -**

* Adeyemi Omisakin
* Limei Zhang
* Omprakash Uppala
* Sathish Reddy

**7th May 2020**

**Prof: Dr. Anol Bhattacherjee**

**Table of Contents**

[Executive Summary](#_TOC_250010) 2

[Problem Significance](#_TOC_250009) 2

[Data Source/Preparation](#_TOC_250008) 2

[Data Sourcing 11](#_TOC_250004)

[Data Cleaning and Preparation 11](#_TOC_250003)

[Hypotheses](#_TOC_250008) 2

[Descriptive Analysis 8](#_TOC_250006)

[Models 8](#_TOC_250006)

[Model 1: 11](#_TOC_250004)

[Model 2: 11](#_TOC_250003)

[Model 3: 11](#_TOC_250003)

[Model 4: 11](#_TOC_250003)

Comparison of Models: 11

[Quality Checks 13](#_TOC_250002)

[Independence check 11](#_TOC_250004)

[Multi-collinearity 11](#_TOC_250003)

[Autocorrelation 11](#_TOC_250003)

[Recommendations 13](#_TOC_250001)

[References 14](#_TOC_250000)

Executive summary: -

After reading about the infant mortality rates in the world and particularly in USA, our interest was driven to find out what factors have a higher influence on survival of infants in USA and to look into any such factors which can have some implementable outcomes to improve infant mortality. To pursue this analysis, the data has been gathered from CDC- Centers for Disease Control and Prevention. The data included the basic birth and death statistics along with many other factors like mother’s age, education, race, weight at the time of birth, place of birth etc. A standard approach has been followed by performing the exploratory analysis, forming hypothesis and conducting the regression analysis on the data. The data has been modelled using 4 models and finding the one which predicts the data best and also is practically suitable.

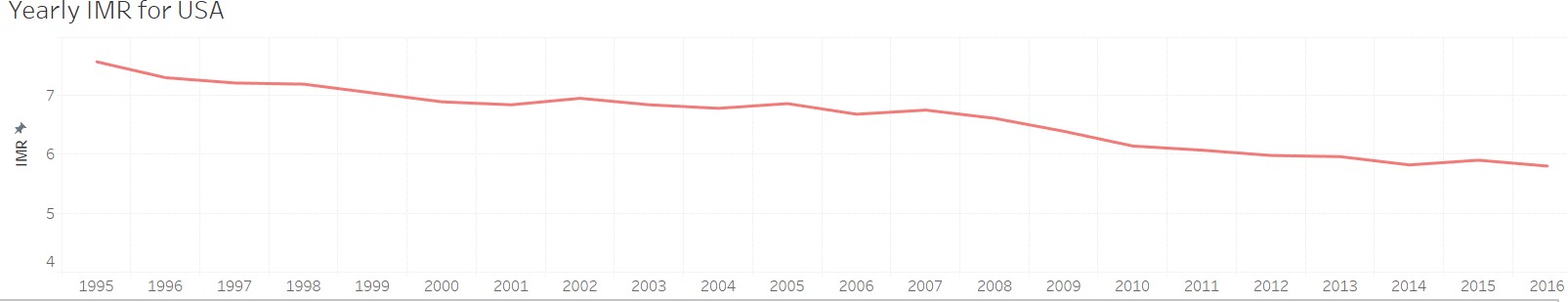
The key findings from our analysis suggest that the age and race of the mother do influence the survival of the infant and the factors like birth weight and delivery method also have an impact on survival of an infant. This report aims at making some actionable recommendations to doctors and practitioners who can make an impact on the patients(to-be-mothers).

Problem Significance: -

Every year in the United States, around 23,000 children die before reaching their first birthday. About half of these children die on the day they are born. But if the baby is African-American, they are more than twice as likely to die. Despite being the largest Economy nation by GDP, U.S has one of the worse Infant Mortality rates in the Industrialized world. This is disheartening. Especially, after knowing that many of these deaths could in fact be preventable. While the doctors perform various medical tests on a regular basis to check the health condition of the Mother and the baby, it is sometimes too late. Interestingly, there are often certain indicators lurking in the data that could help identify the patterns and help prevent many of these deaths.

Family practitioners and Obstetricians often have a huge responsibility in preventing these deaths by identifying these indicators pertaining to the demographic and other characteristics of the mother and/or the child, along with the results from the medical tests. Because, after all the pre-conception and pre-natal cares and all the efforts put in during delivery, it is distressing to see the baby die.

This report aims to identify these trends and list those potential indicators that these doctors could consider in preventing those deaths. The major sources of inspiration for this problem are the trends and the Millennium Development Goals established by the UNICEF itself and the other associated organizations. The MDG pertaining to the child mortality was to reduce the number of deaths by two-thirds, from 87 in every 1,000 births in 1990 to 29 in every 1,000 live births. But unfortunately, this has not been met and the trends vary across the world. The following figure shows the trend for the Infant Mortality Rate (in number of deaths per 1,000 live births) from 1995 to 2016:



Although there seems to be declining trend in the Infant Mortality Rate over the years for USA, there are certain questions to ponder. Is it declining at the rate that we wanted it to? Are the advancements in the technology really contributing to this rate? Does this rate vary across different Ethnicity groups and regions within the country? And so on. While few of the factors like economic disparities and the access to the health care etc., are not in the hands of the practitioners and doctors, the indicators and other factors that could help in identifying and preventing the deaths in mainly discussed in this report.

1. **Data Sourcing and Preparation**
   1. Data Sourcing

To start with any Analysis, it is a known fact that the credibility and reliability of the data collected is of paramount importance. The entire data for this project has been sourced from the official website, Center for Disease control and prevention, CDC, the leading national public health agency of the United States. The staff at CDC collects all the public health surveillance data by working closely with all the state and local health departments, experts from the CDC programs with prevention and control responsibilities and other partner organizations. Together, they develop and implement consistent standards, tools, training and technology to help ensure that disease reporting systems are integrated within each state and local authorities.

* 1. Data Cleaning and Preparation

While majority of the models built, and analysis done are on the data from the raw data files sourced directly from the CDC official public website, there are certain instances where a few trends and graphs will be plotted with the IMR rates (Infant Mortality Rate) instead. However, these are just the aggregate values from the same data taken from the same source.

The data in these files is quite overwhelming and contained millions of observations for the child births information alone. Since, loading and analyzing on these massive datasets at this level is quite challenging, a few random samples have been taken from this gigantic child births data set and collated with the data on the child deaths. The data required for this project has been prepared in the following sequence and approach:

* + 1. The entire child births data file consisted of 11 million plus observations with more than 120 attributes or columns. And the infant deaths data file consisted of around 21,000 observations.
    2. Hence, three random samples – one with the same size as that of deaths, one with twice as many observations and the other with thrice as many observations as that of deaths file have been taken from the births data file.
    3. The files consisted of the variables on the demographic, ethnicity and other mother and child characteristics, along with the type and place of births or delivery. All the redundant columns such as those specifying the descriptions and codes have been removed.
    4. Once the individual files have been obtained, common columns were identified, and the deaths file has been combined with each of these sampled births data files using the SQL queries.
    5. An extra column has been added in these records with two values ‘0’ for deaths and ‘1’ for births to identify

the status respectively. This also forms our Target variable.

* + 1. The variables with the values or categories ‘Unknown’ in majority of the cells have been removed as they do not contribute to any meaningful interpretations.

After going through the papers on various studies conducted by the WHO and UNICEF on the child mortality rates and plotting various distributions and graphs (further discussed in the subsequent sections), the following set of the Predictor and Target or Dependent variables have been selected:

|  |  |
| --- | --- |
| Target Variable Independent Variables | |
| ***Survived (1) / Death (0)*** | Age of the Mother |
| Ethnicity |
| Place of Birth |
| Marital Status |
| Weight of the child at Birth |
| Multiple Births |
| Delivery Method |

1. **Core Hypotheses**

Now that all the data has been cleaning and the necessary variables have been obtained, the core hypotheses can be formulated to know exactly what we are going to test or infer in this project. Again, the basic inspiration behind formulating these hypotheses are the various distributions (discussed in the following section) and the studies conducted by UNICEF and the Millennium development Goals.

The following hypotheses have been laid out for this project:

1. H1a: 𝛽Young < 𝛽Middle-aged> 𝛽Old

The ideal age for a woman to have a baby is between 24 Years – 34 Years. As they get older, the health problems of the mother increase the risk of miscarriage or could affect the health of a child and hence, the probability of survival for the babies of younger women is lesser than that of the middle-aged women. At the very younger age, the body may not be fully ready for the giving birth to a child.

1. H2a: 𝛽Married > 0

Married women have good awareness and are well prepared for pregnancy and take good care of the health when compared to the unmarried women. Hence, the Married women should have higher probability of infant survival when compared to the unmarried women.

1. H3a: 𝛽African-American < 0

The African-American women are known to have worse birth outcomes when compared to the other major ethnic group. This is mainly because of the poverty rates, access to the health care and other inherent health problems that are widely popular among these women.

1. H4a: 𝛽Higher-Education > 0

As the mother is more educated, she is more aware of the issues and better prepared for the pregnancy or the pre-natal care.

1. H5a: 𝛽Less-baby-weight < 0

The babies born with low weight have higher risk of health-related problems and the survival probability is less.

1. H6a: 𝛽Hospital > 0

The probability of survival for the infants of the women giving birth in the hospital are supposed to be higher than that of the rates of women giving birth in the other places. This is obvious as the hospitals have a safe environment and a proper care will be taken by the doctors or the other practitioners.

1. H7a: 𝛽Multiple-births < 0

The twins or triplets often make the women weaker and possess higher risk of deaths when compared to singles. Hence, the probability of survival is less for the multiple births.

1. H8a: 𝛽Cesarean < 0

The probability of survival in cases of natural deliveries are expected to be higher than the cesarean methods since lesser emergency cases or abnormalities are involved with this method when compared to Cesarean method.

1. H9a: 𝛽doctor > 0

The probability of survival for the infants should be higher when the delivery is treated by the professionals like doctor when compared to other medical attendants.

1. H10a: 𝛽Young-mother \* 𝛽Low-Child-Weight < 0

The baby’s health depends on the Mother’s health conditions. Babies having low weight at birth and that are born to young mothers have higher risk of health problems and should be having low survival probability.

1. H10a: 𝛽Young-mother \* 𝛽Unmarried < 0

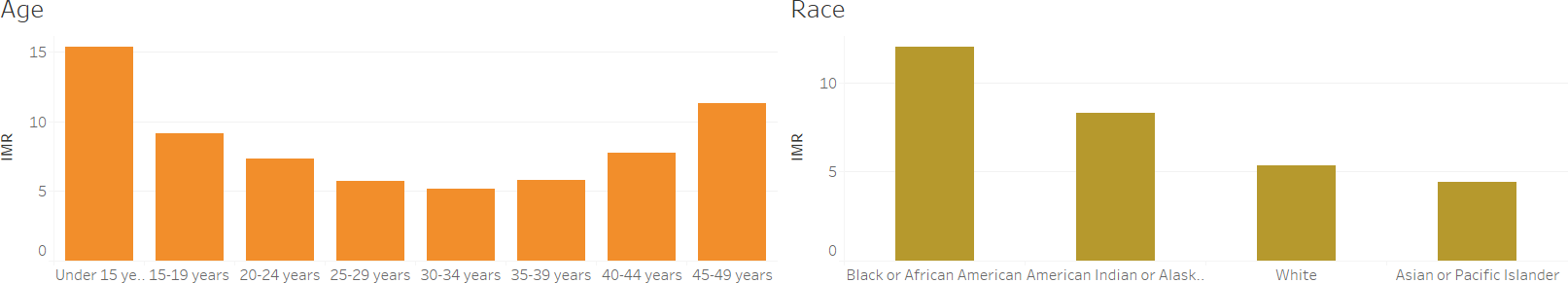
The survival probability for the infants of the young mothers who are unmarried is supposed to be higher since they have less awareness and won’t be prepared to take up the responsibility to take good care of the child and themselves.

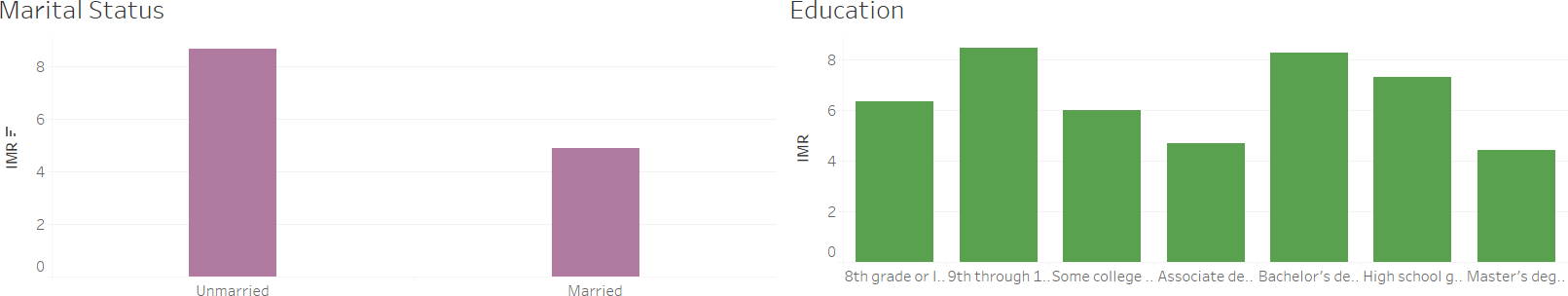
1. **Descriptive Analysis**

The following distributions have been plotted for all these variables or characteristics that are thought to have a potential impact on the mortality rates:

*All the following figures show the distributions of Infant Mortality Rate, IMR (measured as the number of deaths per 1,000 live child births) against various Mother and Child Characteristics.*

* 1. Mother Characteristics





* + 1. Mother’s Age

It is clearly visible from the above distribution of the Mother’s Age that the IMR (Infant Mortality Rate) is lower for the mothers in the age range of 24 – 34 years. There is a distinct pattern for the distribution of the Mother’s Age, supporting our hypothesis and thus making it into our model.

* + 1. Race/ Ethnicity

The African-American women category has the highest Infant Mortality Rate when compared to other Ethnicity groups. This is in line with our hypothesis and various other studies conducted by the United Nations organizations responsible for the Mother and child death control and prevention.

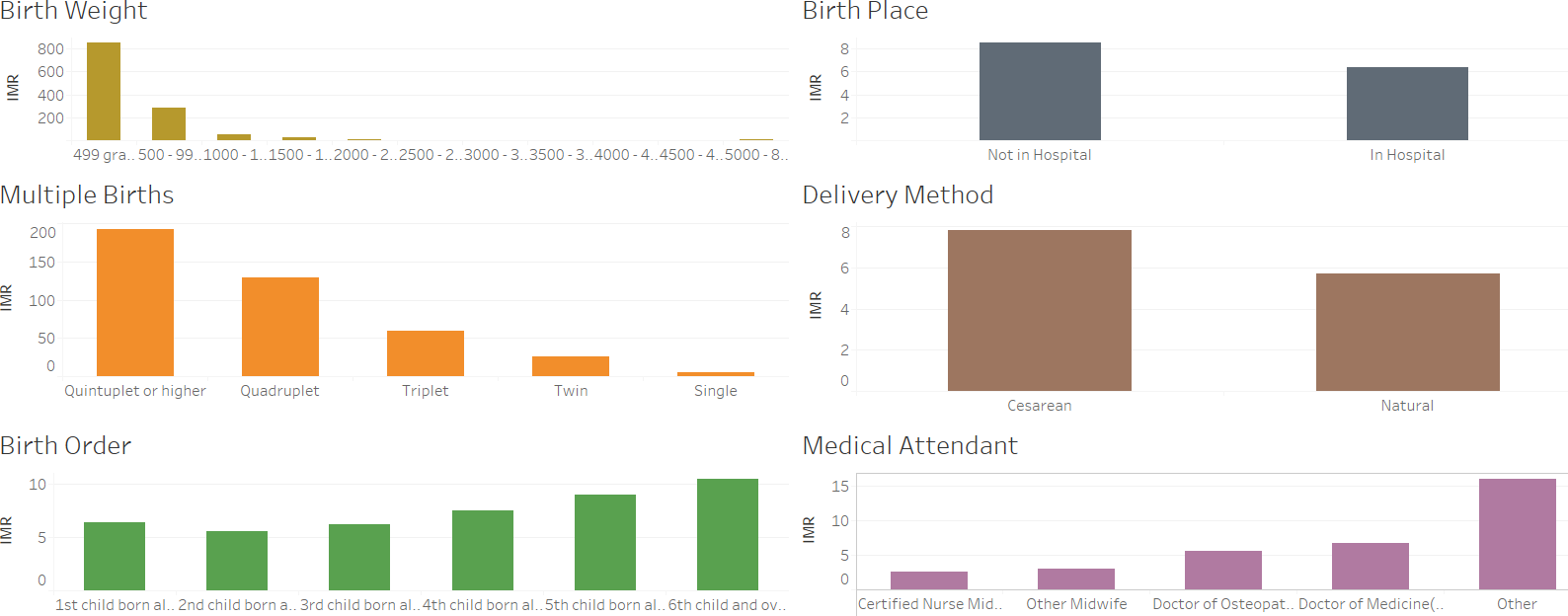
* + 1. Marital status

From the distribution of the IMR vs Marital status plot, it is clear that Unmarried women posses higher risk for the infant deaths when compared to the Married women.

* + 1. Education of the Mother

From the plot of IMR vs Mother’s Education, it appears that the Education of the Mother doesn’t seem to have a major affect on the infant deaths, as opposed to our assumption. Therefore, not included in the model.

* 1. Child Characteristics



* + 1. Birth Weight

The ideal weight for the infants during birth is between 2500 – 4000 grams according to various studies. The babies born with very low weight have the highest risk of deaths which is also apparent from the plot.

* + 1. Birth Place

Hospitals provide the safe environment for the infants during delivery and after birth. Hence, it is not surprising to see the Mortality rate higher for the births at places other than hospitals.

* + 1. Multiple Births

Multiple births such as twins, triplets and quadruplets often make the mother weak, thus increasing the risk of the deaths in such infants.

* + 1. Delivery Method

Cesarean deliveries are often associated with emergencies and abnormalities and hence have a higher rate of IMR, which is also apparent from the plot.

* + 1. Medical Attendant

This plot is counter intuitive and shows higher Mortality rate for the deliveries treated by the doctors compared to other attendants. Therefore, not included in the model.

1. **Modelling**
   1. Model building

Based on the above hypotheses and the descriptive analysis, the following model has been formulated:

*Since the target variable is binary and we want to predict the probability of Survival, the model has been built using logistic Regression and the analysis has been done in R.*

The entire data set has been split into training and testing data sets. 75% of the observations are included in the training data set and the remaining 25% observations are included in the testing data set. The observations in the training data set are used to build the model required to predict the new values. And the observations in the testing data set are used to test the model built using the above approach. Finally, all the models are compared using the Prediction Accuracies and the best model is selected.

* + 1. Model 1

Survival = f (Mother Age, Ethnicity, Place of Birth, Marital status, Weight of the child at Birth,

Multiple Births, Delivery Method, *Mother Age with Child Weight at Birth, Mother Age with Marital Status).*

(The differences in the predictor variables for each model have been highlighted in italic).

* + 1. Model 2

Survival = f (Mother Age, Ethnicity, Place of Birth, Marital status, Weight of the child at Birth, Multiple Births, Delivery Method, *Mother Age with Child Weight at Birth*).

* + 1. Model 3

Survival = f (Mother Age, Ethnicity, Place of Birth, Marital status, Weight of the child at Birth, Multiple Births, Delivery Method, Mother Age with Marital status).

* + 1. Model 4

Survival = f (Mother Age, Ethnicity, Place of Birth, Marital status, Weight of the child at Birth, Multiple Births, Delivery Method).

* 1. Model Comparison

The following metrics have been used to compare all the above four models:

* **Accuracy:** The model accuracy is calculated based on the number of outcomes the model has been able to predict correctly in the test data set consisting of the new observations. And so, the higher the model accuracy the better the model is.
* **Specificity:** Specificity indicates the percentage of Deaths correctly classified or predicted by a model. The higher the Specificity, the better the model will be.
* **Sensitivity:** Sensitivity indicates the number of Survived outcomes that are correctly classified or predicted by a model. The higher the Sensitivity, the better the model.
* **AUC:** It stands for the ‘Area Under the Curve’ and is an estimate of the probability that a classifier or model will rank a randomly chosen positive instance higher than a negative instance. It basically measures the ratio between the hit rates to the false alarms. Hence, the higher the AUC values the better the model or classifier is.

For any test, there usually is a trade-off between the sensitivity and Specificity.

And Since, in this scenario, it is of higher importance to identify a death outcome correctly (Specificity) than the of identifying a survival outcome correctly (Sensitivity), Specificity has been considered as a metric to compare the models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset with equal number of survived and Death outcomes – (Base Accuracy -50%)** | | | |
| **Model** | **Accuracy** | **Specificity** | **AUC** |
| Model 1 | 77.72% | 78.99% | 0.7772 |
| Model 2 | 77.16% | 74.98% | 0.7716 |
| Model 3 | 82.63% | 92.20% | 0.8929 |
| **\***Model 4 | 82.41% | 91.88% | 0.9906 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset with twice as many survived observations than Deaths – (Base Accuracy -66.67%)** | | | |
| **Model** | **Accuracy** | **Specificity** | **AUC** |
| Model 1 | 77.29% | 79.77% | 0.7477 |
| Model 2 | 82.12% | 87.99% | 0.7789 |
| Model 3 | 86.22% | 95.86% | 0.8921 |
| \*Model 4 | 86.24% | 95.56% | 0.8908 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset with thrice as many survived observations than Deaths – (Base Accuracy -75%)** | | | |
| **Model** | **Accuracy** | **Specificity** | **AUC** |
| Model 1 | 76.99% | 80.38% | 0.7366 |
| Model 2 | 86.67% | 92.87% | 0.8058 |
| Model 3 | 88.65% | 97.29% | 0.8891 |
| \*Model 4 | 88.61% | 97.38% | 0.8883 |

8 predictors have been considered for each of the above 4 models. And in addition to these:

1. Model 1 contains two interaction terms Age of the Mother with Birth of the Child and the Age of the Mother with the Marital Status. Accuracy and specificity values and the interaction terms have been observed to be not so significant.
2. Since, the results showed that the interactions are not significantly affecting the outcome, one of these interactions with the lowest significance and effect (Age of the mother with the weight of the child) has been removed in Model 2. Accuracy and the Specificity values slightly improved.
3. Model 3 is also similar to Model 2, but, here the interaction term is (Age of the Mother with the Marital status). The Accuracies and the specificity, AUC values are more or less the same as those of the Model 2.
4. Model 4 is a plain model with all the 8 variables as predictors without any interaction terms. The Accuracy, specificity and AUC values for this model are quite higher when compared to the other three models.

Based on the above analysis and results, the model 4 having the highest Accuracy, Specificity and AUC values has been selected and used for predictions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Hypothesis** | **Accepted/Rejected** | **Significance** | **Inference** |
| H1 | Accepted | Medium | The odds of survival for the infants of middle aged women is higher than that of young and old  women. |
| H2 | Accepted | High | The odds of Survival for the infants of the Married women is higher than that of the Unmarried. |
| H3 | Accepted | High | The odds of Survival for the infants of the Black women is higher than that of the mothers of  other Ethnicities. |
| H4 | Accepted | High | The odds of Survival for the infants having low weight at birth is lower than other infants. |
| H5 | Rejected | - | Place of the birth does not seem to have any major effect on the Survival of the child in U.S. |
| H6 | Rejected | - | Multiple births do not seem to have any major effect on the Survivals. |
| H7 | Accepted | Medium | The odds of survival of the infants when the delivery is of Cesarean type is lower when  compared to natural. |
| H8 | Rejected | - | The interaction between the Age of the mother  and the Weight of the child doesn’t seem to have  a major effect. |
| H9 | Rejected | - | The interaction between the Age of the mother and the Marital status doesn’t seem to have a  major effect. |

1. **Quality Checks**

In addition to the high prediction accuracies and the Specificity and AUC values, the model has been built using the logistic Regression approach, the following assumptions need to be satisfied to ensure the quality of the Analysis:

* 1. Independence check

Since all the observations in the data set are independent and the outcome of any particular child is not dependent on the outcome of any other child, this assumption has been satisfied for this analysis.

## Multi-collinearity

All the variables or predictors in the model are categorical and are not dependent on each other and no variable can be predicted from any of the other variables. Hence, this assumption has also been satisfied.

## Auto-correlation

All the analysis has been performed on the data in the year 2015 and hence there is no chance of Auto- correlation in this analysis.

1. **Recommendations**

From our analysis, we may make few recommendations to the health practitioners and doctors which can be implemented to improve the survival rate among the infants.

* From the results of our analysis, we see that the mother’s age is indeed a crucial factor for child birth. Doctors and practitioners can educate their young patients (young mothers) to take time and revive their decision to continue with pregnancy, as there is always a risk of lower emotional maturity among young mothers and also their knowledge about care and precautions to be taken during the pregnancy.
* From the statistical results, it is evident that the infants of African American race are at a greater risk of mortality among other races. So, the doctors need to take a closer look at the ethnicity details of patients and try to prevent any known health risks of that race and educate them about the preventive care that should be taken care of.
* One of the well-known fact is further established from our analysis is the importance of birth weight of the baby at the time of the birth. Though doctors advise patients to take proper diet during pregnancy, there is still a need to reinforce this more efficiently. With the growing fashion of maintaining low calorie diet may hinder the nutrition supply to the fetus which may ultimately lead to a low birth weight and putting the baby at a risk of infant mortality.
* In addition to the above-mentioned recommendations, it is equally important to focus on the delivery methods adopted by the patients. Educating and increasing awareness in this area can help prevent infant deaths at the time of birth due to unavailability of pediatric/ gynecological support to the baby and mother. The doctors should strive to make the delivery in a natural way rather than cesarean. Cesarean should be adopted only in demanding situations, where there is no option to go ahead with natural way of birth.

These are some of the recommendations which we think would be incremental towards improving the survival rates among infants in USA. The onus is not only on the policy-makers but also on the doctors and practitioners who have the first-hand experience of dealing with the patients (the mothers to be). Their directed efforts on some of the specific areas can have a large impact in improving the survival rate. We wish every effort should be made to save the infant from any preventable risk.