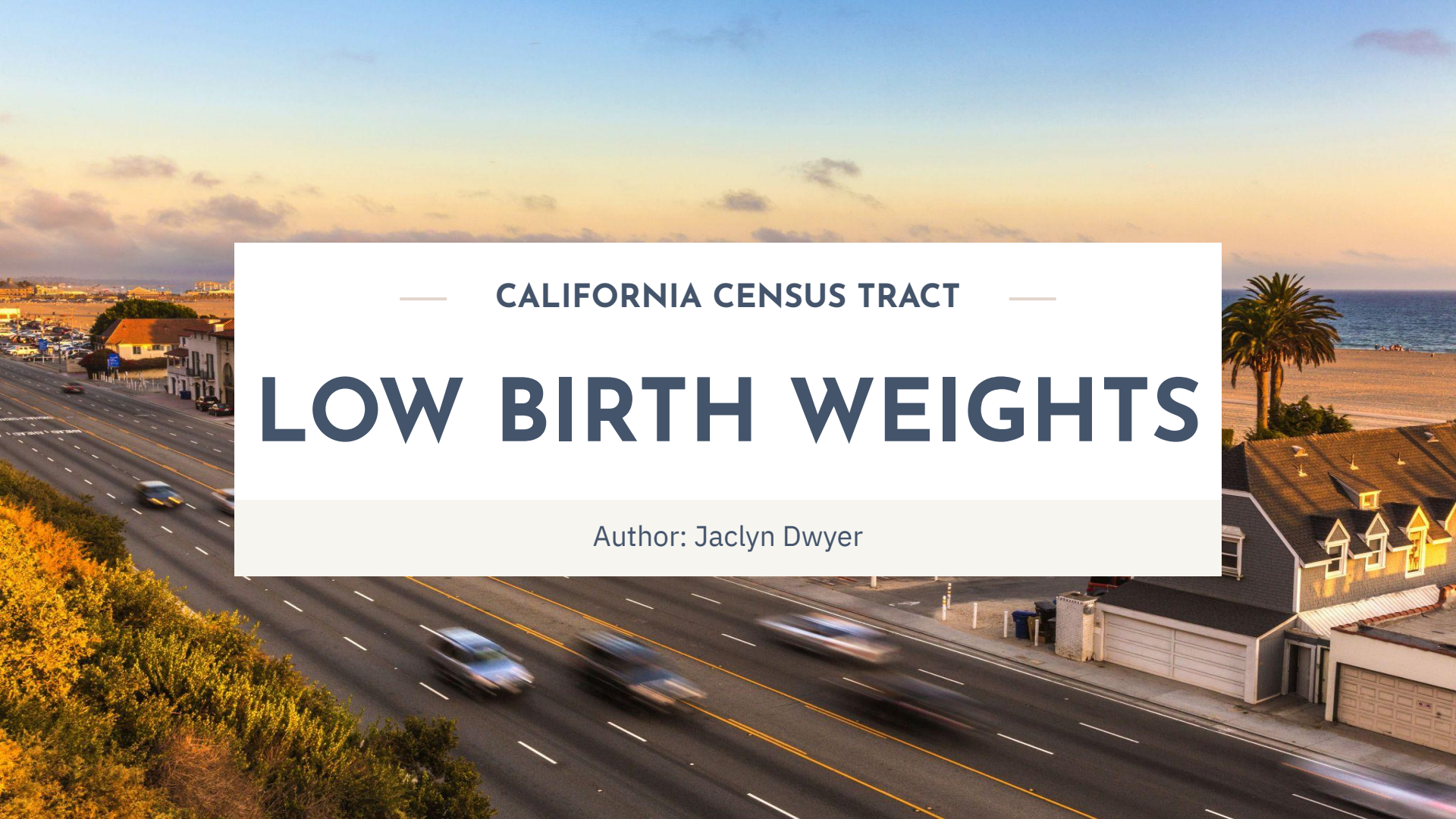


CALIFORNIA CENSUS TRACT

LOW BIRTH WEIGHTS

Author: Jaclyn Dwyer

BUSINESS PROBLEM

The goal of this project is to predict the percentage of low birth weight (LBW) births in California census tracts based off their population characteristics and environmental health hazards to help Kaiser Permanente determine how to identify areas that need higher level NICUs.

LOW BIRTH WEIGHT (LBW)

- Born early or restricted growth
- Babies born less than 5.5 lbs
  - Neonatal Intensive Care Unit (NICU)
  - Low oxygen levels
  - Feeding tubes
- Nervous system problems

Why population characteristics?

Certain population characteristics have been associated with increased LBWs including certain ethnicities.

Why environmental health hazards?

Studies suggest environmental health hazards have been linked to increased risks for LBWs.

DATA PROCESSING

California Communities Environmental Health Screening Tool reports (CalEnviroScreen CES) released by the Office of Environmental Health Hazard Assessment (OEHHA) - aim to identify California census tracts that are burdened and vulnerable to multiple pollution sources.

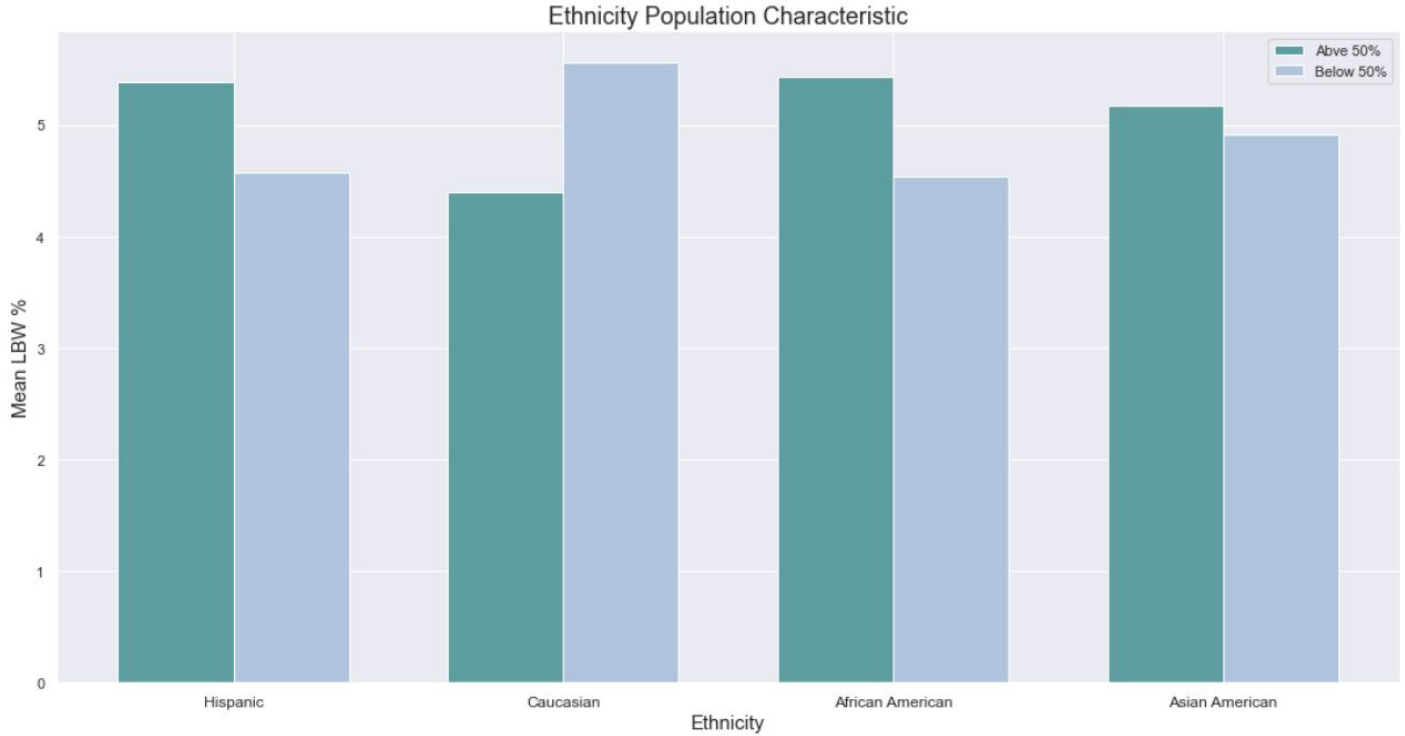
- CES 3.0 Published 2018
- 8035 California census tracts
- 56 Columns
- Dropped columns and rows :
  - Not pertinent to business problem
  - Zero total population
  - Missing LBWs ~211
- Added previous environment information from CES 2.0 report
- Added demographic information
- Added smoking prevalence

CalEnviroScreen

OEHHA

California Office of Environmental Health Hazard Assessment

VISUALIZATIONS - POPULATION CHARACTERISTICS



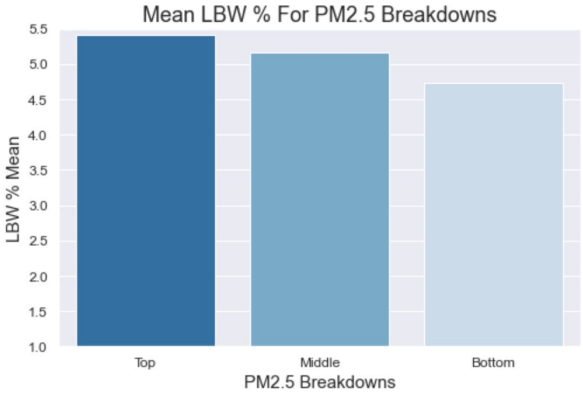
KEY TAKEAWAYS:

LBW % ↑

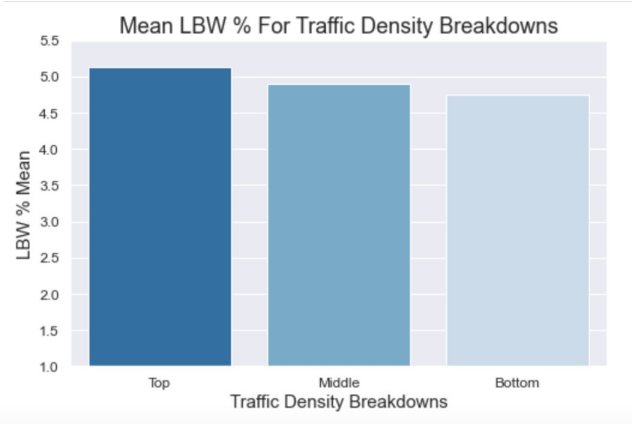
- *More* Hispanics
- *Less* Caucasians
- *More* African Americans
- *More* Asian Americans

VISUALIZATIONS - ENVIRONMENTAL HEALTH HAZARDS

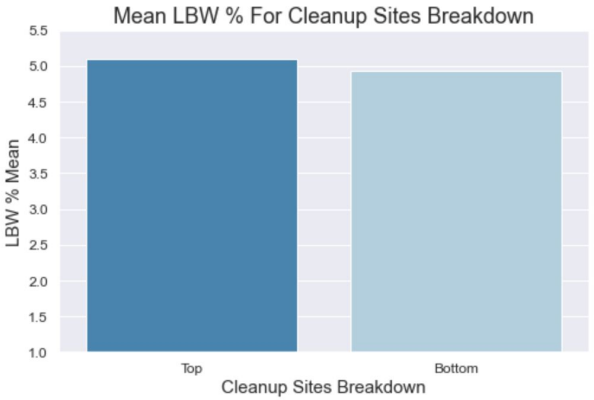
Studies linked increase LBWs with exposure to particulate matter (PM 2.5), traffic, and cleanup sites.



Higher mean LBW % with more PM exposure



Higher mean LBW % with more traffic exposure



Higher mean LBW % with more cleanup site exposure

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

MODELS

Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

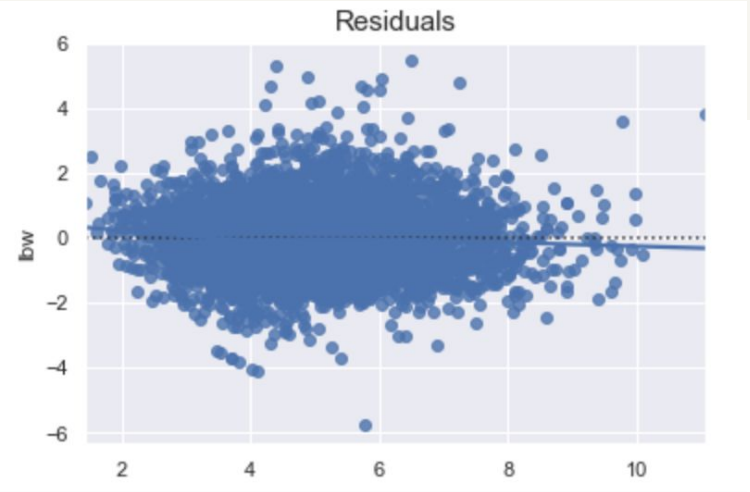
MODELS

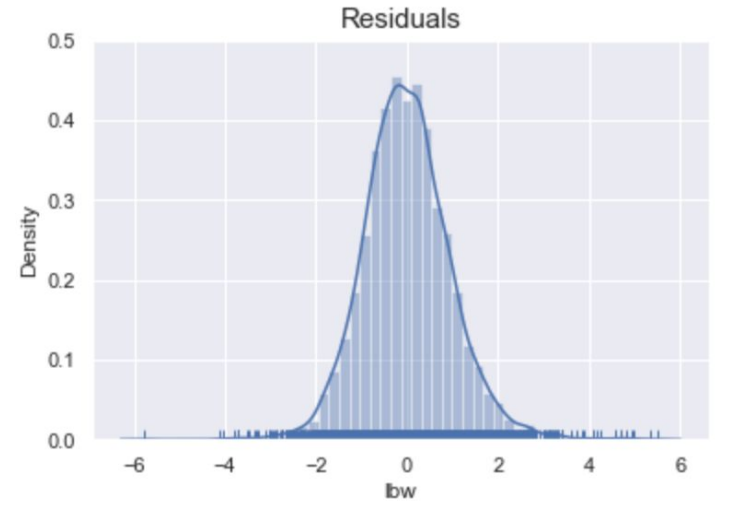
Evaluation Metrics: RMSE Score & R-Squared Score (>0.5)

Model	Train RMSE	Test RMSE	Train R-Squared	Test R-Squared
Baseline:	0.9927	1.0018	0.5946	0.568
Interactions:	0.8294	1.1163	0.7169	0.4596
Kbest:	0.9655	0.9742	0.6164	0.5884
RFE:	0.9657	0.9746	0.6163	0.5881
GridSearch Random Forest:	0.9320	1.0754	0.6426	0.4985

FINAL MODEL

- About:
  - Linear Regression Model
  - 98 Selected Features with RFE
  - 30 Significant for predictions
- Scoring:
  - RMSE ~ 0.97
    - Model is off by about 0.06 on average compared to the entire range of the target variable
  - R-Squared
    - 62% of variance in train set
    - 58% of variance in test set
- Assumptions:
  - Homoscedasticity
  - Normality





education, white, prev_lbw,
african_american_breakdown_More,
total_population_and_african_american,
total_population_and_disadvantaged_Yes,
total_population_and_white_breakdown_More,
ozone_and_white,
ozone_and_african_american,
ozone_and_prev_lbw,
ozone_and_disadvantaged_Yes,
pm2_5_and_prev_lbw,
education_and_yrs_11_64,
linguistic_isolation_and_prev_lbw,
housing_burden_and_hispanic,
housing_burden_and_white_breakdown_More,
less_10_yrs_and_prev_lbw,
less_10_yrs_and_disadvantaged_Yes,
less_10_yrs_and_hispanic_breakdown_More,
yrs_11_64_and_white,
yrs_11_64_and_disadvantaged_Yes
greater_65_and_white,
hispanic_and_african_american,
hispanic_and_disadvantaged_Yes,
white_and_prev_lbw,
african_american_and_prev_lbw,
african_american_and_african_americ
an_breakdown_More,
other_and_white_breakdown_More
prev_lbw_and_hispanic_breakdown_Mo
prev_lbw_and_african_american_break
down_More

education, white, prev_lbw,
african_american_breakdown_More,
total_population_and_african_american,
total_population_and_disadvantaged_Yes,
total_population_and_white_breakdown_More,
ozone_and_white,
ozone_and_african_american,
ozone_and_prev_lbw,
ozone_and_disadvantaged_Yes,
pm2_5_and_prev_lbw,
education_and_yrs_11_64,
linguistic_isolation_and_prev_lbw,
housing_burden_and_hispanic,
housing_burden_and_white_breakdown_More,
less_10_yrs_and_prev_lbw,
less_10_yrs_and_disadvantaged_Yes,
less_10_yrs_and_hispanic_breakdown_More,
yrs_11_64_and_white,
yrs_11_64_and_disadvantaged_Yes
greater_65_and_white,
hispanic_and_african_american,
hispanic_and_disadvantaged_Yes,
white_and_prev_lbw,
african_american_and_prev_lbw,
african_american_and_african_americ
an_breakdown_More,
other_and_white_breakdown_More
prev_lbw_and_hispanic_breakdown_Mo
prev_lbw_and_african_american_break
down_More

NEXT STEPS

Some future steps to improve this project include:

- Adding in clustering to help with predictions
- Running gradient tree boosting and XGBoost models to see if better predictions are achieved
- Creating a causal inference model

THANK YOU!

github.com/jhdwyer

linkedin.com/in/jaclyn-henn-dwyer

jaclyndwyer.medium.com