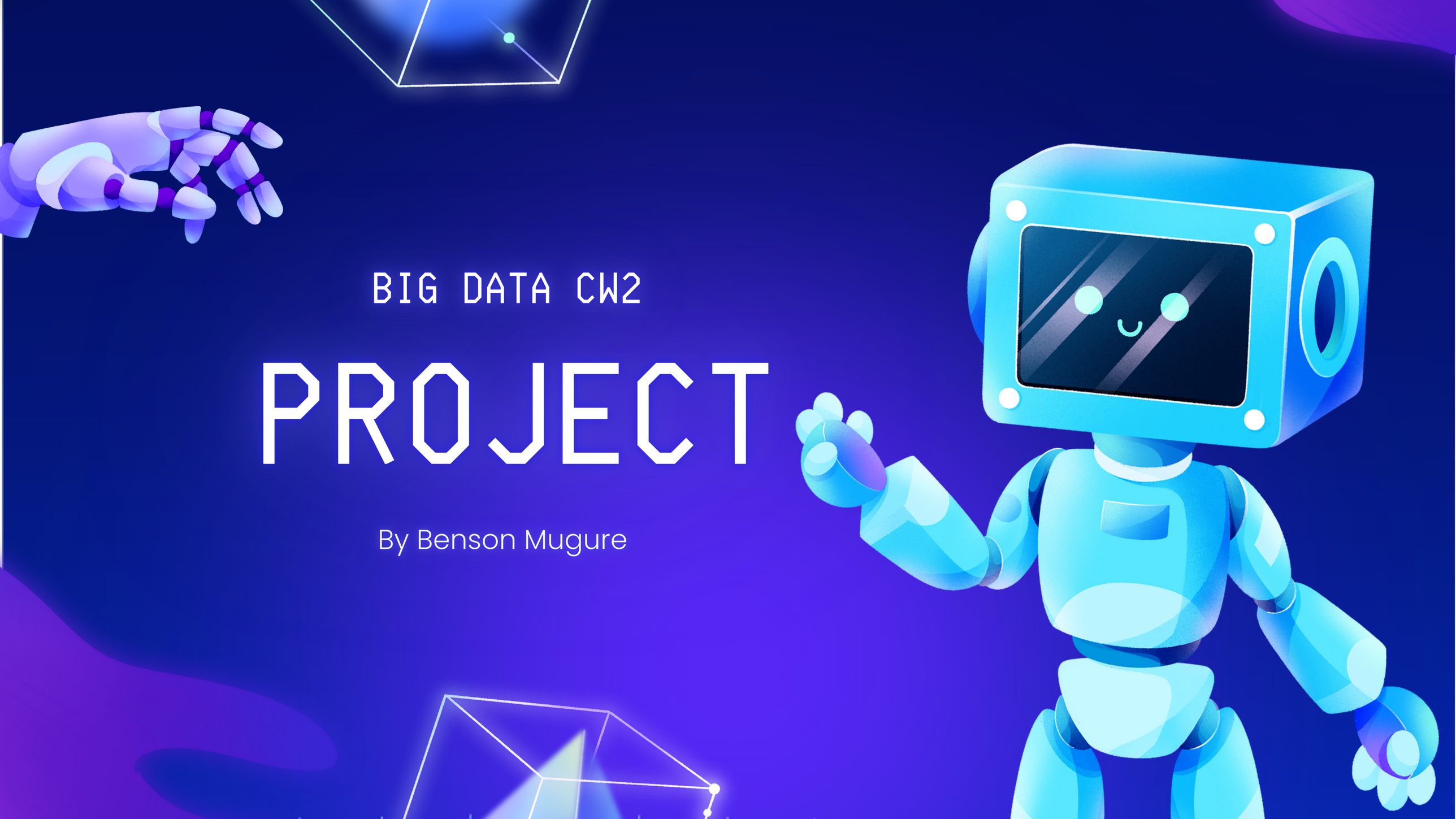
BIG DATA CW2

PROJECT

By Benson Mugure



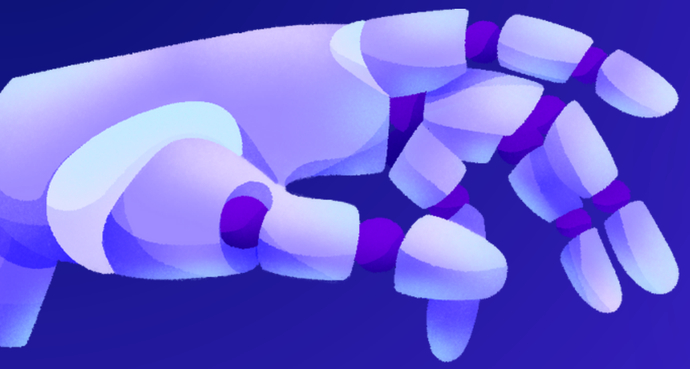
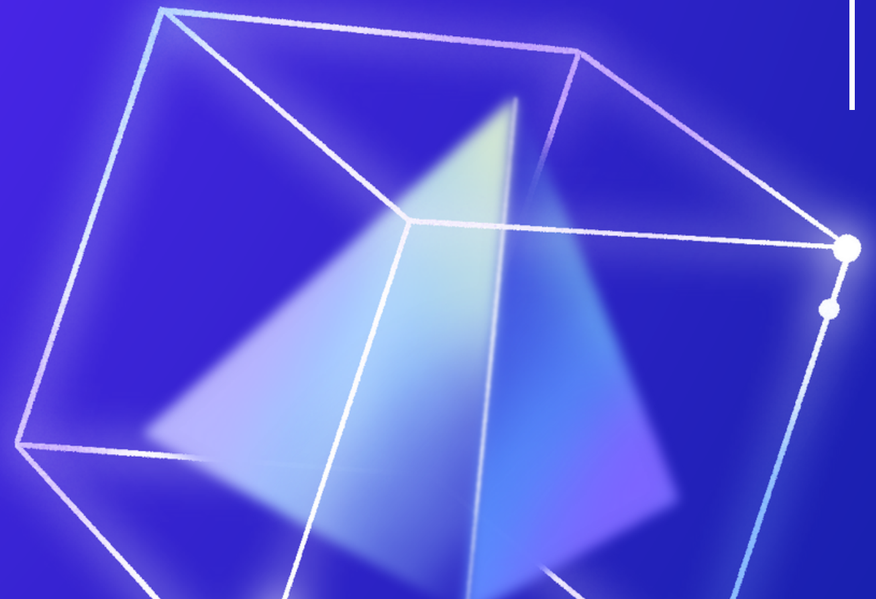


TABLE OF CONTENTS



PROBLEM OVERVIEW

The domain of shipping and logistics plays an integral role in global trade and commerce, serving as a critical conduit for the movement of goods across borders and continents. The accurate estimation of shipping times represents a pivotal challenge faced by stakeholders in this industry.

Solution: Machine Learning Predictions

This coursework embarks on a meticulous exploration of predictive models aimed at estimating shipping times, dissecting the complexities, challenges, and nuances entrenched within the logistics domain. Spanning data exploration, model construction, and evaluation, this endeavor navigates through multifaceted facets in a bid to unravel the optimal strategies for forecasting shipping durations.

MY DATASET

- I found my dataset from a closed Kaggle competition.
- It had 14 columns and 5114 columns
- The columns include a timestamp one as well as several categorical ones describing shipping details

EDA AND PRE-PROCESSING

01

Uniqueness, correlation and categorical data:

1. I dropped the columns with only 1 unique value
2. I dropped the features that were highly corrected according to my heatmap
3. I encoded categorical variables into numeric ones

02

Variance:
My boxplot figure showed that the figures in the gross_weight feature were much bigger than in the other columns.
I scaled it as well as other columns so that they have values between -1 and 1

03

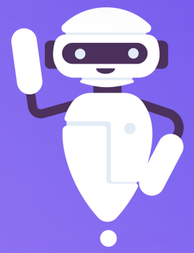
Feature Engineering:
Dealing with timestamp:

I divided the timestamp into day, month and year then applied cyclical encoding (sine and cosine transformations) for cyclical patterns to avoid my model from interpreting them as regular numbers

04

Dealing with Null values:
My dataset had no null values

MODEL CONSTRUCTION & TUNING



SVM MODEL

I used the SVR implementation of SVM available in sklearn library since my problem was a regression one.

I used automated methods, such as GridSearchCV and RandomSearchCV, to find the best parameters (C and gamma)

NEURAL NETWORK

I used Keras from Tensorflow to make a Neural Network with 3 layers.

For tuning, I adjusted my number of layers, adjusted the model architecture to add a drop off as well as using optimizers and learning rate scheduling

OTHER MODELS

I tested out 8 other models which could be applied to a regression problem and the best one was a bayesian ridge regression model

MODEL EVALUATION

01

My best SVM model had the following metrics:

- MSE: 48.86
- RMSE: 6.99
- MAE: 3.96
- R-squared: 0.55

02

My best Neural Network had the following metrics:

- MSE: 49.577
- RMSE: 7.041
- MAE: 3.971
- R-squared: 0.55

ANALYSIS

My SVM model performed better than my neural network. This was to be expected given that I have a small dataset. Neural Networks work better with larger, less structured datasets

The Bayesian ridge model surpassed my SVM model. It utilizes Bayesian statistics to estimate the coefficients of the regression model. Instead of estimating a single point value for the coefficients, Bayesian Ridge estimates a distribution of possible values for the coefficients. This distribution captures uncertainty in the parameter estimates.

THANK YOU!

APPENDIX

1. Link to dataset source: https://www.kaggle.com/datasets/salil007/1-shipping-optimization-challenge
2. Link to Github: https://github.com/Virgo-Alpha/Shipping_Optimization
3. Link to Notebook: https://gist.github.com/Virgo-Alpha/ba83255854bef25c8b945e8568a6a63f
4. Link to pdf notebook:https://drive.google.com/file/d/1I7Vy133dSJdWDR-kfZyliaJ8PYsqcpVa/view?usp=sharing
5. Link to EDA page: https://virgo-alpha.github.io/Shipping_Optimization/
6. Link to report: https://drive.google.com/file/d/1IX27JJqQSQXahjoB3mT0g7T6OF3xA9uv/view?usp=sharing