Influenza Vaccination Prediction Mode

By

Katta Manasa	20221CSE0685
Harshavardhan Kurthkoti	20221IST0101
Keerthi M Naik	20221IST0103
Aashna K Nishad	20221IST0073

Department of Computer Science Engineering

Presidency University

Itgalpura Rajanakunte,Yehlanka

Bengaluru – 560093 , Ph No – 080 23093500

February 2024

Influenza Vaccination Prediction Model Report

Introduction

This report presents the development and evaluation of a machine learning model aimed at predicting the likelihood of individuals receiving influenza vaccines. With the resurgence of interest in vaccination amidst the COVID-19 pandemic, understanding the determinants of vaccine acceptance and uptake is critical for public health initiatives. By leveraging demographic, behavioural, and attitudinal features collected from the 2009 National H1N1 Flu Survey (NFHS) conducted by the Centers for Disease Control and Prevention (CDC), this study seeks to uncover insights into vaccine hesitancy and inform targeted intervention strategies.

Data Overview

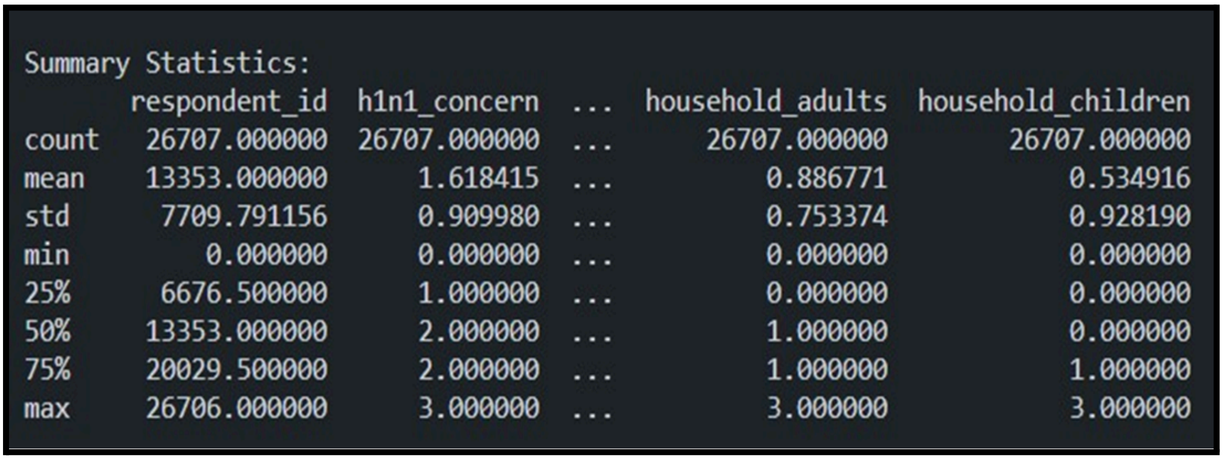
- Training Data:
    - The training dataset comprises 26,707 samples, each characterised by 38 features spanning demographic attributes (e.g., age, education), behavioural indicators (e.g., hand hygiene practices, face mask usage), and attitudinal factors (e.g., perceived vaccine effectiveness, risk perception). This rich dataset provides a comprehensive snapshot of individual-level factors influencing vaccine decision-making.
- Test Data:
    - The test dataset mirrors the structure of the training data, consisting of 26,708 samples for model evaluation. The inclusion of a separate test set ensures unbiased estimation of model performance on unseen data, facilitating robust generalisation to real-world scenarios.

Data Preprocessing

- Handling Missing Values:
    - Missing values were meticulously addressed using a combination of imputation techniques tailored to the nature of the missingness. Mode imputation was applied to categorical features, preserving the distributional properties of the data, while median imputation was utilised for numerical features to mitigate the impact of outliers.
- Feature Engineering:
    - Feature engineering played a pivotal role in transforming raw data into meaningful representations suitable for model training. Categorical features underwent one-hot encoding to capture non-ordinal relationships, while the 'age_group' feature was discretized to replace age ranges with the average age within each group, enhancing interpretability without sacrificing information.

Exploratory Data Analysis (EDA)

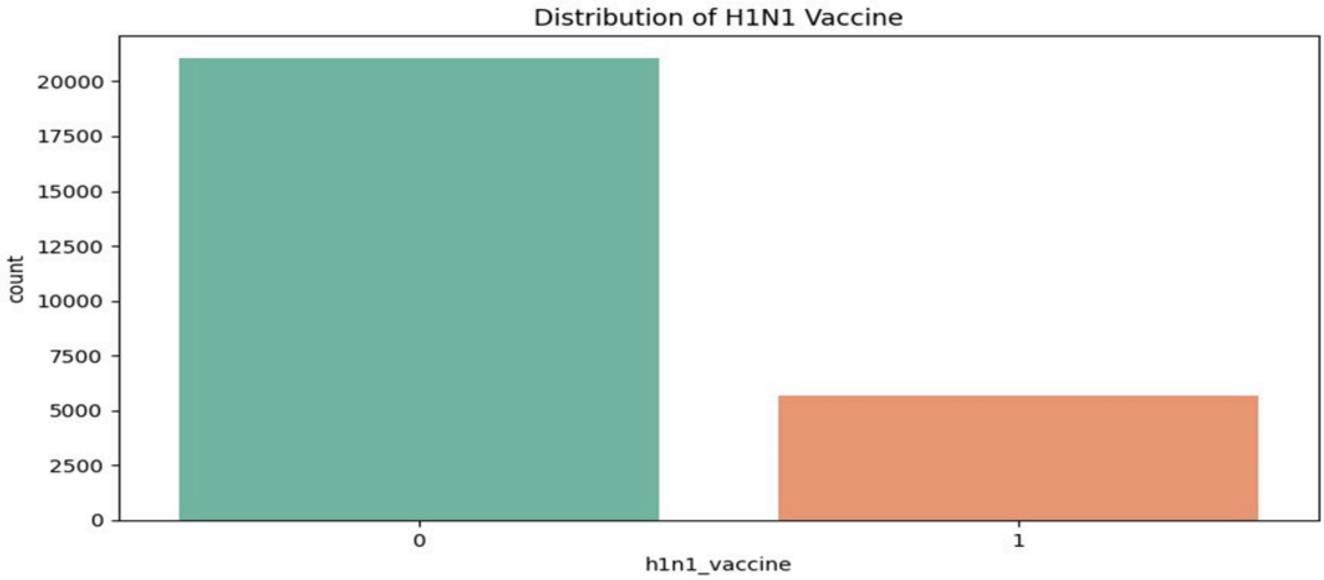
- Importance of EDA:
    - EDA served as the cornerstone of this study, enabling a deep dive into the underlying structure and patterns of the dataset. Through visualisations and statistical summaries, EDA unveiled hidden insights and guided subsequent modelling decisions, ensuring a holistic understanding of the data.
- Summary Statistics:
    - Summary statistics provided a bird's-eye view of key dataset characteristics, shedding light on central tendency, dispersion, and outliers. These insights facilitated hypothesis generation and guided feature selection strategies, laying the foundation for robust model development.



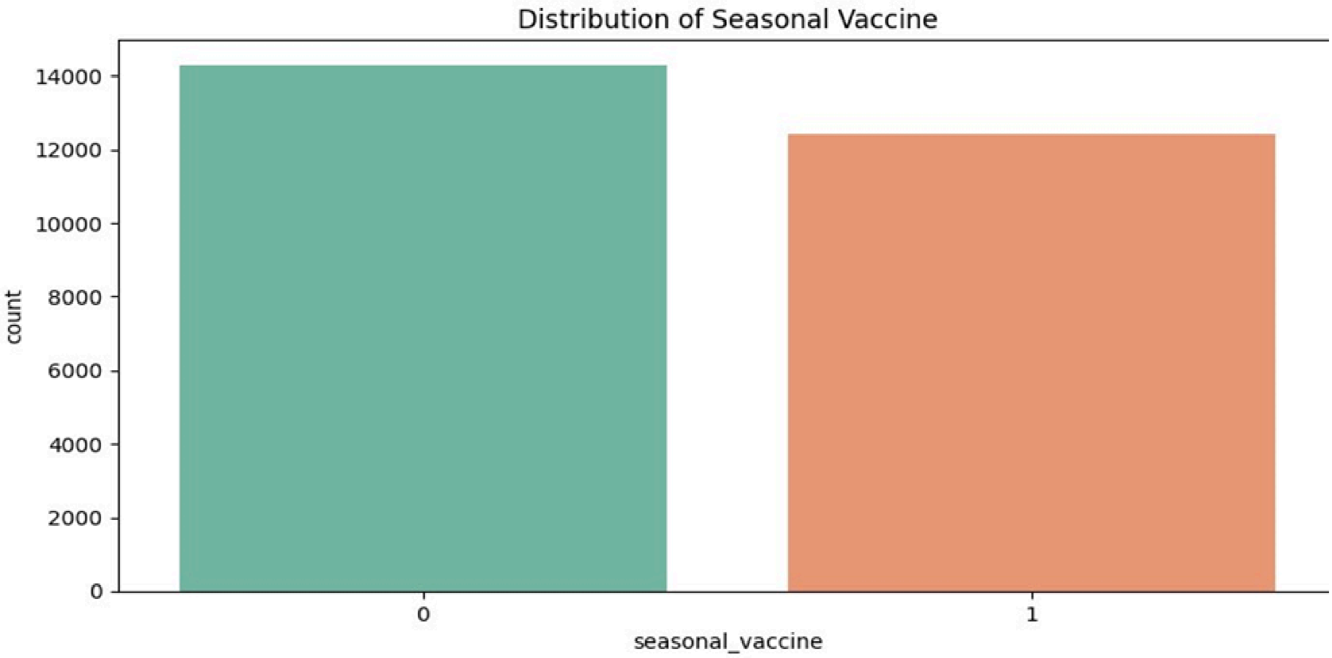
```
Summary Statistics:
       respondent_id  h1n1_concern  ...  household_adults  household_children
count   26707.000000  26707.000000  ...      26707.000000        26707.000000
mean    13353.000000      1.618415  ...          0.886771            0.534916
std      7709.791156      0.909980  ...          0.753374            0.928190
min         0.000000      0.000000  ...          0.000000            0.000000
25%      6676.500000      1.000000  ...          0.000000            0.000000
50%     13353.000000      2.000000  ...          1.000000            0.000000
75%     20029.500000      2.000000  ...          1.000000            1.000000
max     26706.000000      3.000000  ...          3.000000            3.000000
```

- Distribution Graphs
    - Visualisation of vaccination distributions across various demographic and behavioural dimensions unearthed nuanced trends and disparities in vaccine uptake. By stratifying vaccine coverage by demographic subgroups, EDA identified vulnerable populations requiring targeted interventions, thereby addressing equity concerns.

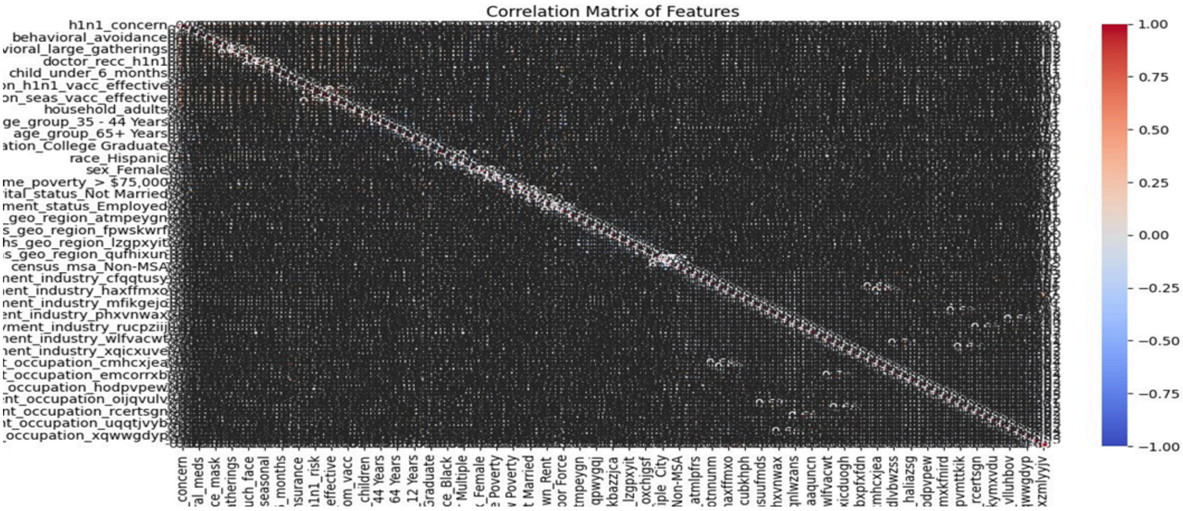
Distribution of H1N1 Vaccine:



Distribution of Seasonal Vaccine:



- Missing Values Heatmap:
    - The missing values heatmap visualises the spatial distribution of missingness across features, offering valuable cues for imputation strategies and highlighting potential biases introduced by missing data. This diagnostic tool enabled data completeness assessment and informed judicious handling of missing values.
- Correlation Matrix:
    - The correlation matrix unveiled inter-feature relationships, illuminating underlying dependencies and redundancies within the dataset. Strong correlations hinted at collinearity, necessitating feature selection or regularisation techniques to mitigate multicollinearity-induced model instability.



Model Development

- Linear Regression Model:
    - A linear regression framework was selected as the modelling paradigm, leveraging its simplicity, interpretability, and computational efficiency. By formulating separate models for predicting H1N1 and seasonal flu vaccination probabilities, the study aimed to capture distinct determinants of vaccine uptake specific to each influenza strain.

Model Training

- Training Split:
    - The training data underwent an 80-20 split into training and validation sets to facilitate model training and evaluation. This partitioning strategy ensured model performance estimation on unseen data, guarding against overfitting and promoting model generalisation.

- Model Training:
    - The linear regression models were trained using state-of-the-art optimization algorithms, fine-tuning model parameters to minimise prediction error and maximise predictive accuracy.
- Model Evaluation:
    - Model performance was rigorously evaluated using the Receiver Operating Characteristic Area Under the Curve (ROC AUC) metric, a gold standard for binary classification tasks. The achieved ROC AUC score of 0.8447 underscored the models' discriminative power and predictive efficacy across both target variables.

Predictions

- Prediction Process:
    - Leveraging the trained models, predictions were generated for the test data, estimating the probabilities of individuals receiving H1N1 and seasonal flu vaccines. These predictions provided actionable insights into vaccination behaviour, empowering healthcare authorities to prioritise resources and interventions effectively.
- Output File:
    - Predictions were seamlessly exported in CSV format, featuring respondent IDs and corresponding vaccination probabilities for both H1N1 and seasonal flu. This standardised format facilitated downstream analyses and decision-making processes.

Conclusion

- Model Performance:
    - The developed linear regression models showcased robust performance in predicting influenza vaccination probabilities, underscoring the efficacy of demographic, behavioural, and attitudinal features in capturing vaccine decision-making dynamics.

- Future Considerations:
    - Looking ahead, further refinement of feature engineering techniques and exploration of advanced modelling architectures hold promise for enhancing predictive accuracy and model

generalisation. Additionally, integrating real-time data streams and leveraging ensemble methods could bolster model robustness in dynamic epidemiological landscapes.

Recommendations

- Utilisation of Predictions:
    - The generated predictions represent a valuable resource for healthcare authorities and policymakers, enabling proactive identification of individuals at heightened risk of vaccine hesitancy. By tailoring intervention strategies to high-risk populations, public health initiatives can optimise resource allocation and maximise vaccine uptake.

- Targeted Intervention:
    - Targeted intervention programs should be designed to address the unique needs and concerns of vulnerable subpopulations identified by the model. By leveraging community partnerships and culturally competent outreach efforts, these interventions can foster trust, reduce vaccine hesitancy, and promote equitable vaccine access.

Limitations

- Data Limitations:
    - Despite the richness of the dataset, certain demographic and attitudinal features were underrepresented, limiting the model's predictive granularity. Future data collection efforts should prioritise comprehensive feature coverage to enhance model performance and robustness.

- Model Assumptions:
    - Linear regression models inherently assume linear relationships between features and target variables, which may not fully capture the complexity of vaccine decision-making dynamics.

Exploring nonlinear modelling techniques and ensemble methods could mitigate this limitation and yield more nuanced insights.

Future Work

- Feature Engineering:
    - Continued exploration of feature engineering techniques, including interaction terms and polynomial features, holds potential for uncovering latent relationships and enhancing model interpretability.
- Model Selection:
    - Experimentation with advanced modelling architectures, such as gradient boosting and neural networks, may unlock additional predictive power and enable the modelling of intricate interactions within the data.
- External Data Integration:
    - Integration of external datasets, such as social determinants of health and geospatial information, can enrich the modelling framework and provide contextual insights into vaccine decision-making behaviours.

References

- Documentation on Linear Regression - Scikit-learn
- Evaluation Metrics for Classification - Scikit-learn