Cervical Cancer Risk Factors

Kristie Kooken

Today’s presentation

- Overview of cervical cancer
- Project Goals: Groundbreaking invention
- Model development
- Project results
- Next steps

Cervical Cancer

- United States → Annually: 11,500 diagnoses & 4,000 deaths
- Globally → 2020: 604,000 diagnoses & 342,000 deaths

- Human papillomavirus (HPV) is thought to be responsible for 90% of all cervical cancers
- HPV is contracted by having sex with a person who has the virus

- Two characteristics of this cancer
  - Women > 30 years old
  - Long lasting infection with certain types of HPV

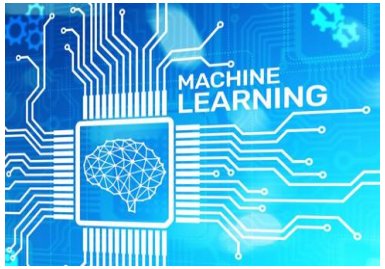
Cervical Cancer

- 93% of all cases of cervical cancer are preventable
  - With screening tests and HPV vaccination

Project Goals

- Develop an invention that uses machine learning
  - Cost effective app-based tool

Health & Screening info



  - Predicting cervical cancer and HPV based on health info & risk factors
  - Gives pathway for next steps for medical care or intervention
  - Ensures privacy

Model Development: Data Source

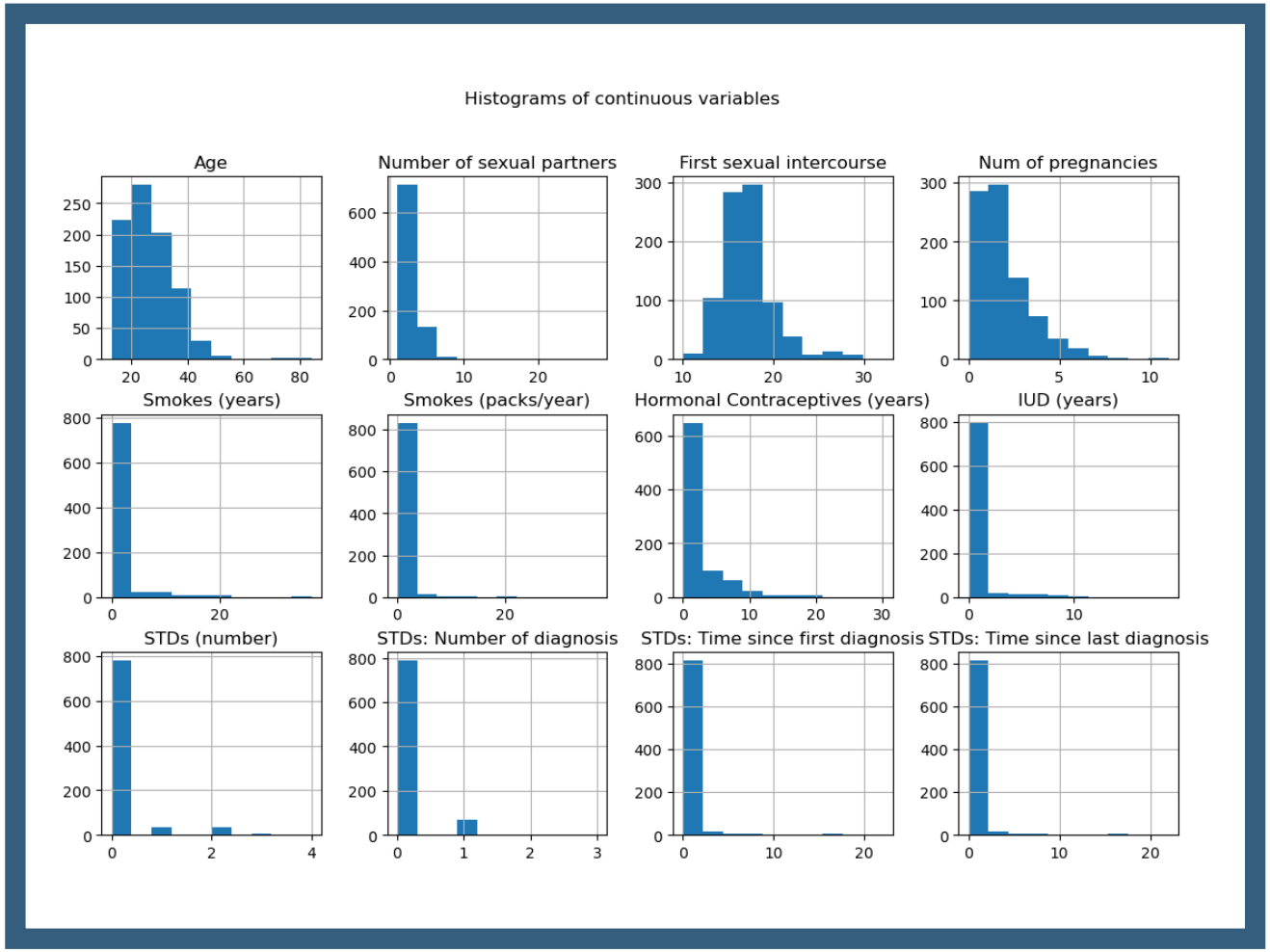
- University of California, Irvine under the name Cervical cancer (Risk Factors) Data Set
- Collected at Hospital Universitario de Caracas in Caracas, Venezuela
- 858 subjects and 36 variables

Model Development: Variables

- Binary categories:
  - Cancer diagnosis, HPV diagnosis
  - STD Y/N questions
  - Smoking Y/N
  - Birth Control Y/N
- Continuous:
  - Age
  - Length of STDs
  - Length of Smoking
  - Length of Birth Control

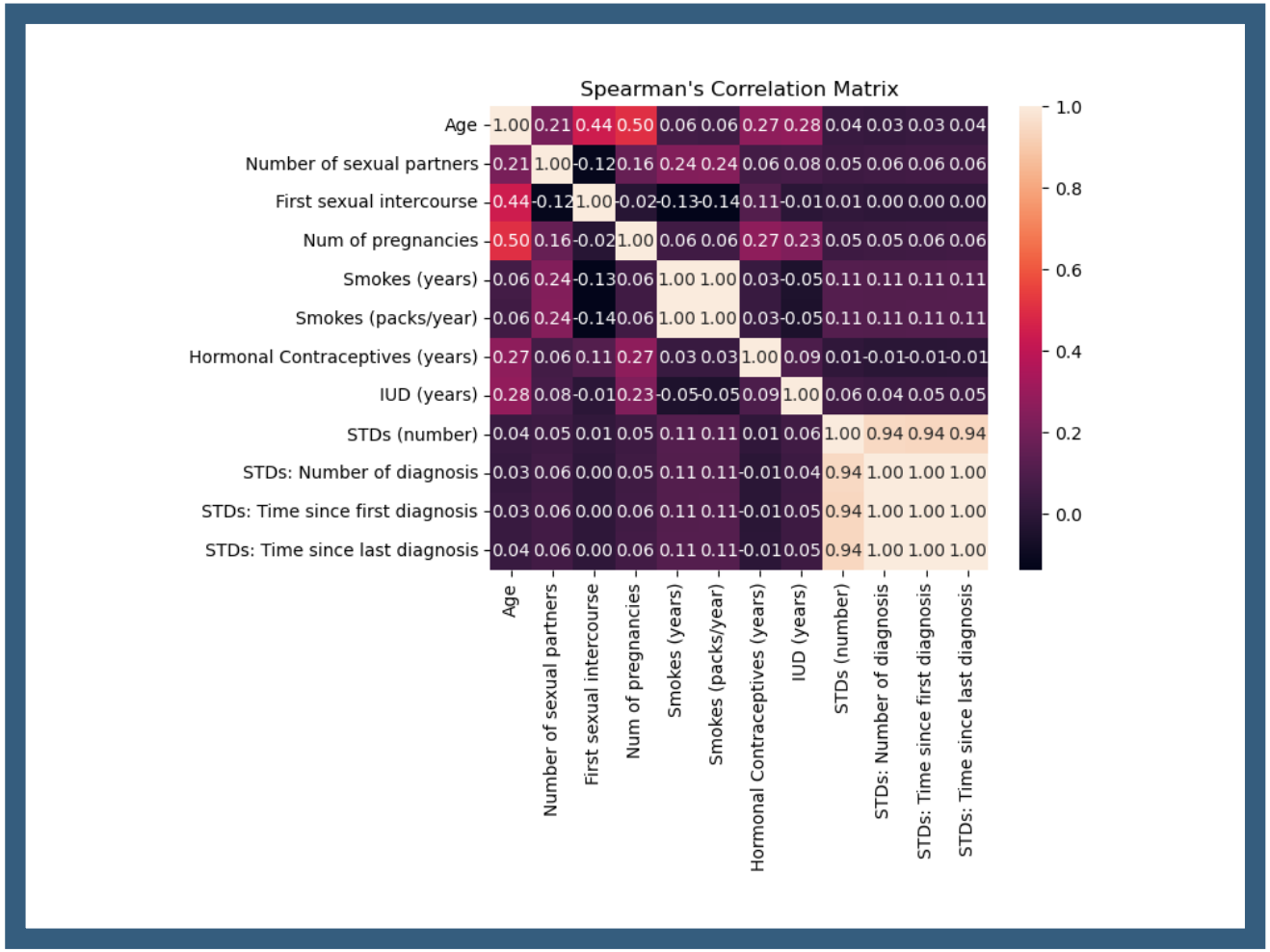
Model Development: Understanding the data

- All females
- Average age 26.8 years
- Cervical cancer diagnoses make up 2% of the data
  - 18 cases
- HPV diagnoses make up 2% of the data
  - 18 cases
- 16 out of 18 cervical cancer & HPV diagnoses are for the same patients



Model Development: Data cleaning

- Replace missing values
- Keep all records due to rare outcome
- Check distributions before and after missing value imputation



Model Development: Dropping variables

- Using correlation to determine relationship between continuous variables
- 4 variables are dropped
- +2 variables with no documentation

Model development: Methodology

Logistic regression: Standardized input

- 2 models, Cervical Cancer, HPV

Feature Selection

- Keeping only relevant features

Addressing imbalance

- Diagnoses of cervical cancer or HPV is rarer

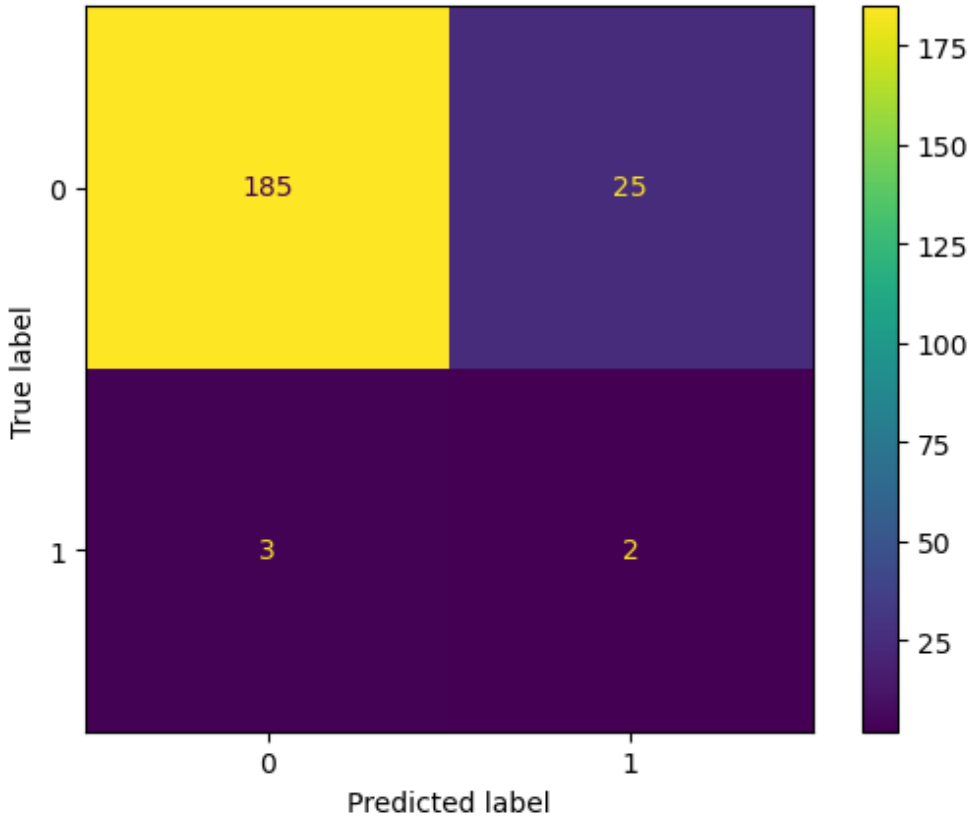
Final model on balanced data

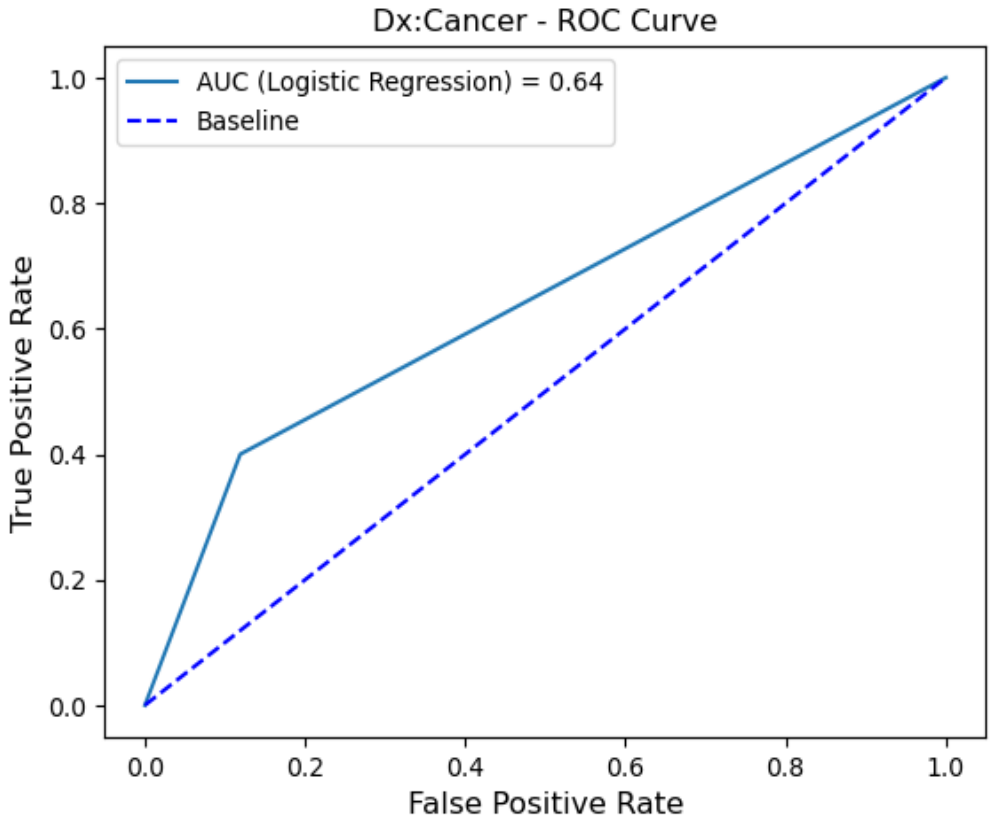
- Performance metrics

Results: Logistic Regression final model

Model	Accuracy	Class	Precision	Recall	F1	Support
Baseline Logistic Regression	0.98	0	0.98	1.00	0.99	210
	-	1	0.00	0.00	0.00	5
Final Logistic Regression	0.87	0	0.98	0.88	0.93	210
		1	0.07	0.40	0.12	5

Model Results: More accurate but shows room for improvement





Conclusions

The ability to accurately predict a cervical cancer or HPV diagnosis would be a great advantage for all humanity.

Privacy concerns will remain a concern.

- Information gained from this effort does demonstrate a need for further modeling
  - Ask:
    - More funding
    - Subject matter expert
    - More time

Questions ?

References

HPV Vaccination: What Everyone Should Know (Accessed 2023, August 11). Vaccines and Preventable Diseases. https://www.cdc.gov/vaccines/vpd/hpv/public/index.html

What Should I Know About Screening? (Accessed 2023, August 11). Gynecologic Cancers. https://www.cdc.gov/cancer/cervical/basic_info/screening.htm

HPV-Associated Cancer Statistics (Accessed 2023, August 11). HPV and Cancer. https://www.cdc.gov/cancer/hpv/statistics/index.htm