Predicting Lethal Outcome for Myocardial Infarction Patients

Jacob Weissman

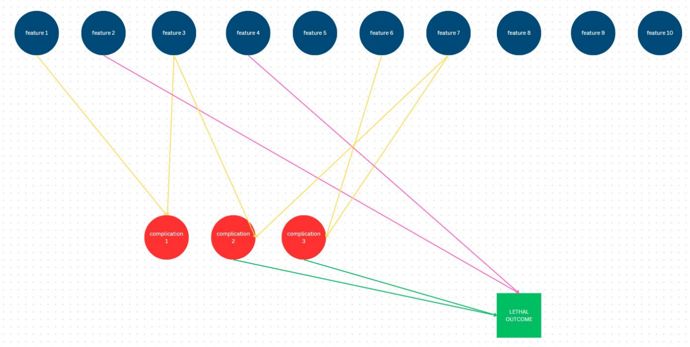
About the Dataset

- Myocardial infarction
- Features + Complications
- Lethal outcome: was cause of death myocardial infarction?
- Imbalance: 84.06% alive
    - stratified split

A	B	C	D	E	F	G	H	I	J
ID	AGE	SEX	INF_ANAM	STENOK_AN	FK_STENOK	IBS_POST	IBS_NASL	GB	SIM_GIPERT
1428	53	0	0	1	1	1		2	0
1430	56	0	0	0	0	0		2	0
1476	79	1	0	6		2		2	0
1481	82	0	1			2		3	0
1485	83	1	1	0	0	2		2	0
1500	64	0	0	5	3	2		0	0
1517	43	1	0	3	2	2		0	0
1525	66	1	0	6	2	0		2	0
1528	61	0	0	5	2			2	0
1529	61	1	3	6	3	2		3	0
1534	70	0	0		2	2		2	0
1538	62	1	0	0	0	2		0	0
1545	60	1	1	0	0	2		3	0
1564	56	1	3	6	2	1		2	0
1580	57	1	0						
1588	65	0	2			1		2	0
1592	41	1	0	0	0	2		0	0
1599	56	1	3	6	2	2		2	0
1605	64	0	0						0
1612	53	1	0	5	2	2		0	1
1621	44	1	0	0	0	2		0	0
1625	84	0	0	6	2	1		2	0

Project Goals

- Predict lethal outcome on **day 3**
  - medical specialists can craft **care plans** and appropriately **monitor** patients to mitigate **risk of death**
- Scoring: **F1**
- Model type: **random forest**
- Incorporate complications data - **available during training ONLY**

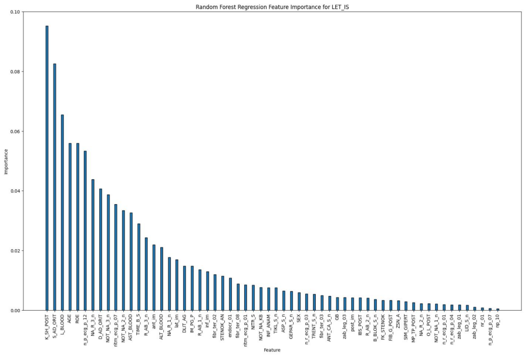


Preprocessing

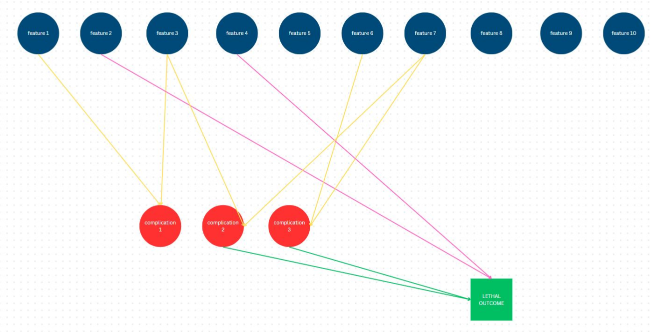
- Avoid data leakage!
- Convert lethal outcome to **binary** dead/alive
- Missing: compare **single-median** and **KNN** imputation
- **Robust** scaling for continuous numerical features

Feature Selection

- Perform separately for **KNN** and **single-median** imputations
- Perform for target **'lethal outcome'** variable (pink arrows)
- Perform for each **complication** (yellow arrows)

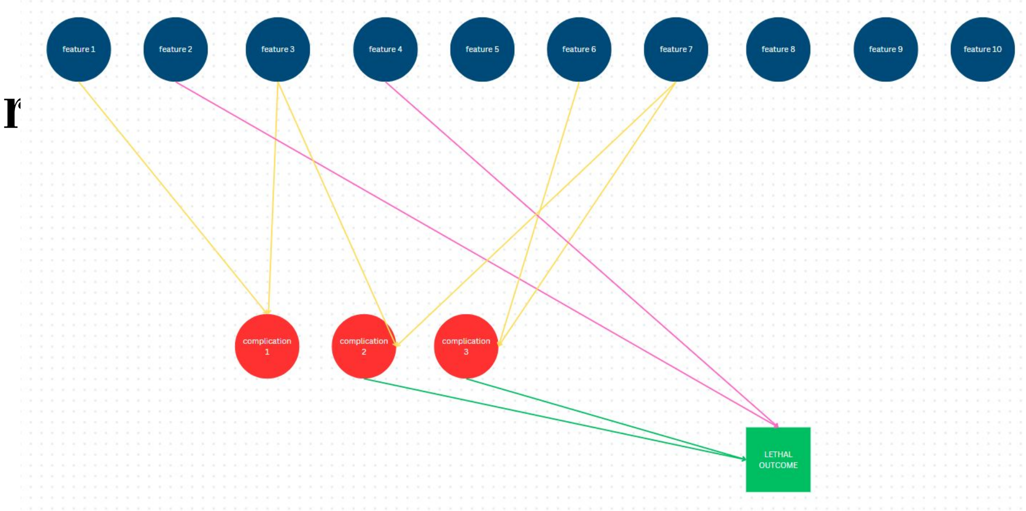


example model's random forest feature importance - elbow method was then used for each target/complication model



Complication Feature Selection

- Perform **complication featur** **selection** (green arrows)
    - select only models with **strong performance** on holdout dataset

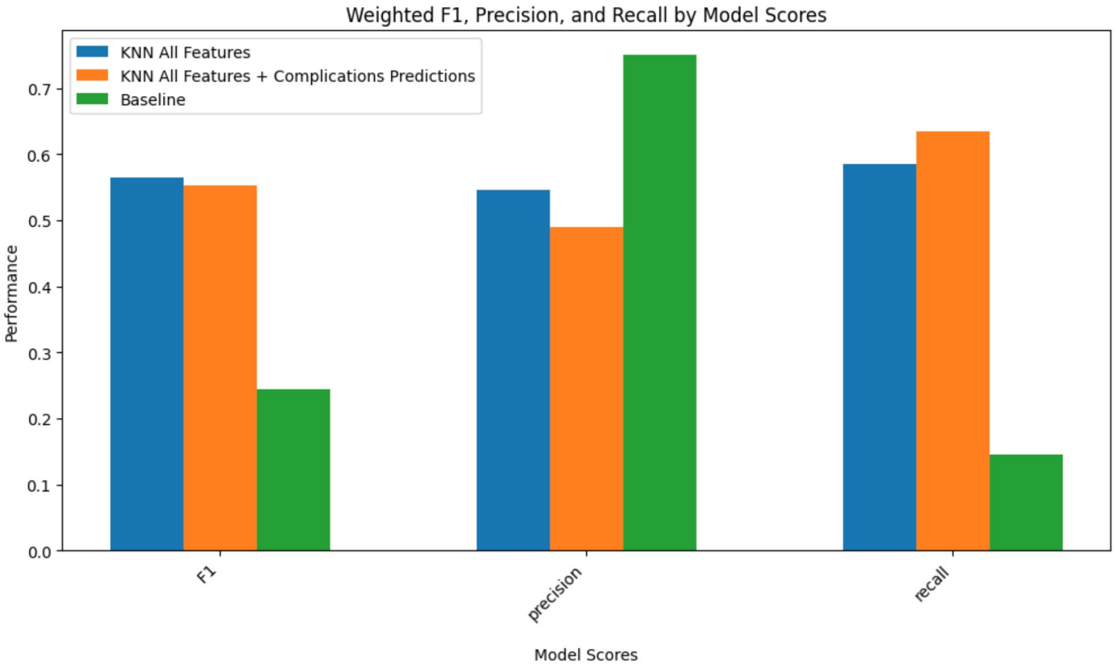


Final Results: Lethal Outcome

KNN All Features: highest performing on a validation dataset, **without predictions**

KNN All Features + Complications: highest performing on a validation dataset, **with complication predictions**

Baseline: single median imputation, no other preprocessing



(test dataset, full training)

Conclusions

- Feature selection unsuccessful
  - highest performing on validation and used for testing -> **all features**
  - too few samples -> too much noise?
  - **small amounts** of information?
- **KNN > single-median**
  - missing at random?
- Poor complication performance
  - too few samples? ambiguous diagnoses?
- **Predicting complications improved recall, is useful**

Thank You!

Jacob Weissman