Maccabi Home Task – DS position

Predictive model to rank patients developing a hypertensive disorder

Simple (EDA)

Dataset Overview

•**10,000 rows × 143 columns** (142 numeric + 1 text).

•**Target column:** y.

•**Prevalence:** ~4.3% positive cases (imbalanced dataset).

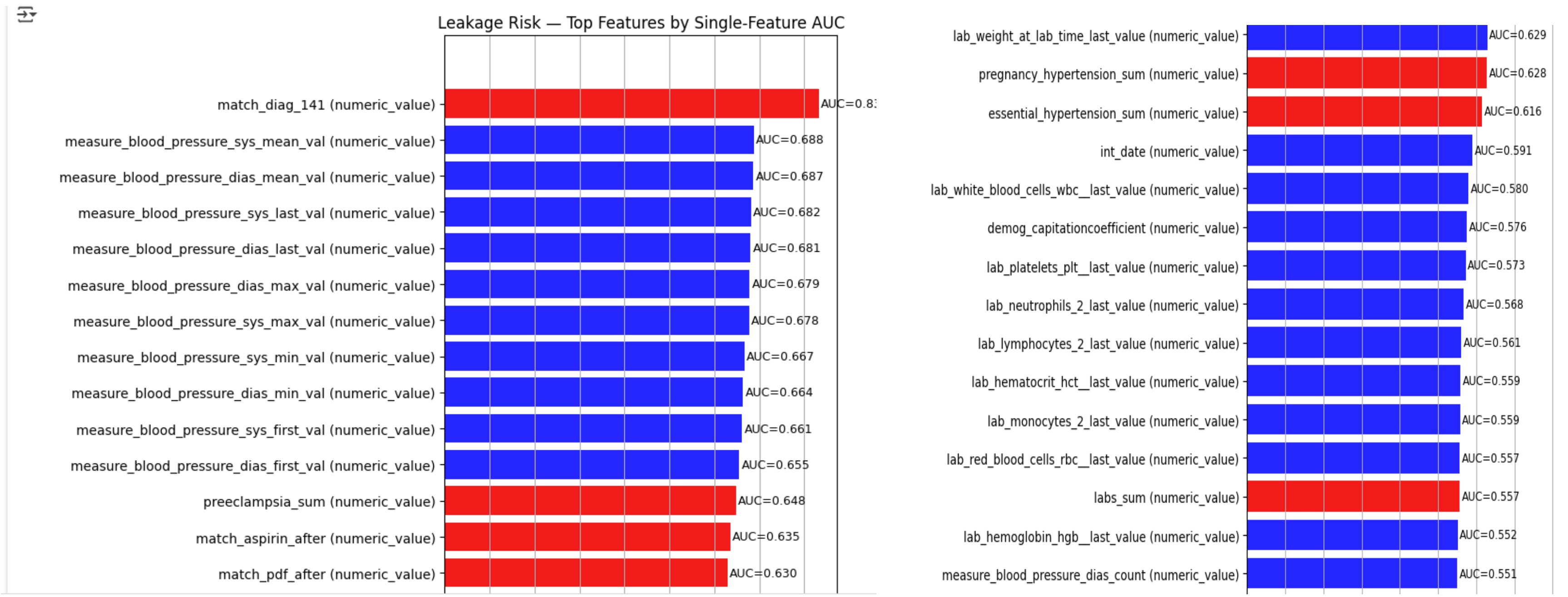
Missing Data

•**57 columns (>10% missing values)** — important to handle before modeling.

•Some columns with **>99% missing**: dropped as raw features, but retained as **is_missing flags** to capture clinical decision signal.

•Missingness itself may indicate **clinical suspicion** (e.g., certain labs ordered only when risk suspected).

Leakage Risk Check — Single-Feature



Early detector for **data leakage**. Any single feature with very high AUC likely encodes post-prediction information and must be excluded for a week-15 model.

match_diag_141 has **AUC ≈ 0.832** (red bar) + . _*sum variables suspected as data leakage columns.

Conclusion:

•Treat match_diag_141 (and any similar match_* and _*sum variables tied to later diagnosis/registries) as **leakage → drop from training**.

•Keep BP features—strong, plausible, and **clinically valid** predictors at week 15.

Text (EDA)

•**Clinical notes column:** “clinical_sheet”.

•Parsed into **7 canonical sections**: complaints, risk factors, findings, labs/imaging, medications, vitals, recommendations.

•**Stopwords removed:** Hebrew fillers (“מתלוננת“ ,”לא“ ,”כן“ ,”ללא”...) and section headers.

•**Sentence splitting**

•**Top n-grams per section** (positives only):

	ngram	count	canon
0	עייפות מוגברת	216	complaints
1	כאבי ראש	166	complaints
2	עייפות מוגברת ובחילות	86	complaints
3	מוגברת ובחילות	86	complaints
4	תחושת עייפות	84	complaints
5	ראש קלים	84	complaints
6	כאבי ראש קלים	76	complaints
7	שיפור בבחילות	73	complaints
8	מדי פעם	73	complaints
9	עייפות מתמשכת	65	complaints

Table 1:1-3 grams for “complication” section

•**Keyword checks: found strong clinical notes:**

" סיכון קל לרעלת היריון עקב תוצאות "

Feature Engineering (tabular features)

•**Lab normalization:** build function to correct inconsistent units as needed for laboratory Results columns.

•**Binary flags:** created *low/high* abnormality indicators based on known limits for laboratory tests (HGB, HCT, WBC, PLT, etc.).

Exploration for *Risk Factors* and creation a new features based on them:

•**Age risk:** added flag for **age > 38 years**.

•**Derived ratios:**

•**NLR = neutrophils / lymphocytes**

•**PLR = platelets / lymphocytes**

•**Blood pressure:** derived **pulse pressure** and **MAP** (mean arterial pressure).

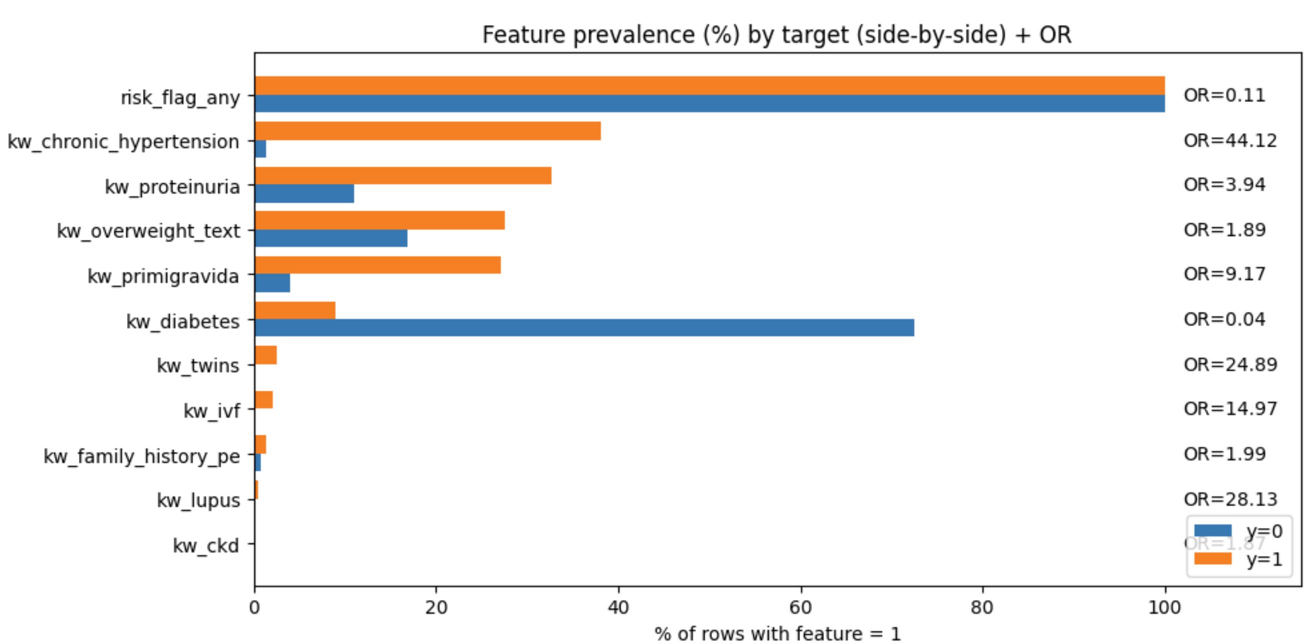
•**History aggregates:** counts of diagnoses in **last 4m vs 24m**, ratio 4/24, flag for “new vs chronic”.

NLP Feature Extraction

•**Regex-based feature extraction:** Built Hebrew keyword lexicon for key risk factors (e.g., diabetes, IVF, family history).

•**Feature engineering:** Created binary indicators (kw_*) and earliest-week variables (kw_*_week) per patient.

•**Aggregate signals:** Derived summary flags such as *any risk factor present* and *minimum week of risk factor appearance*.



Model Evaluation

1. Tabular only

•**ROC-AUC = 0.636, PR-AUC = 0.087**

•That's only a bit better than random (baseline PR ≈ prevalence ~0.04).

2. Text only (complications + risk factor section)

•**ROC-AUC = 0.972, PR-AUC = 0.696**

•Extremely strong! This means that the way complications/risk factors are documented carries **very predictive signal**.

•Risk: still may contain *semi-leaky phrasing* (doctors often write things very close to diagnosis).

3. Fused (tab + text)

•**ROC-AUC = 0.896, PR-AUC = 0.521**

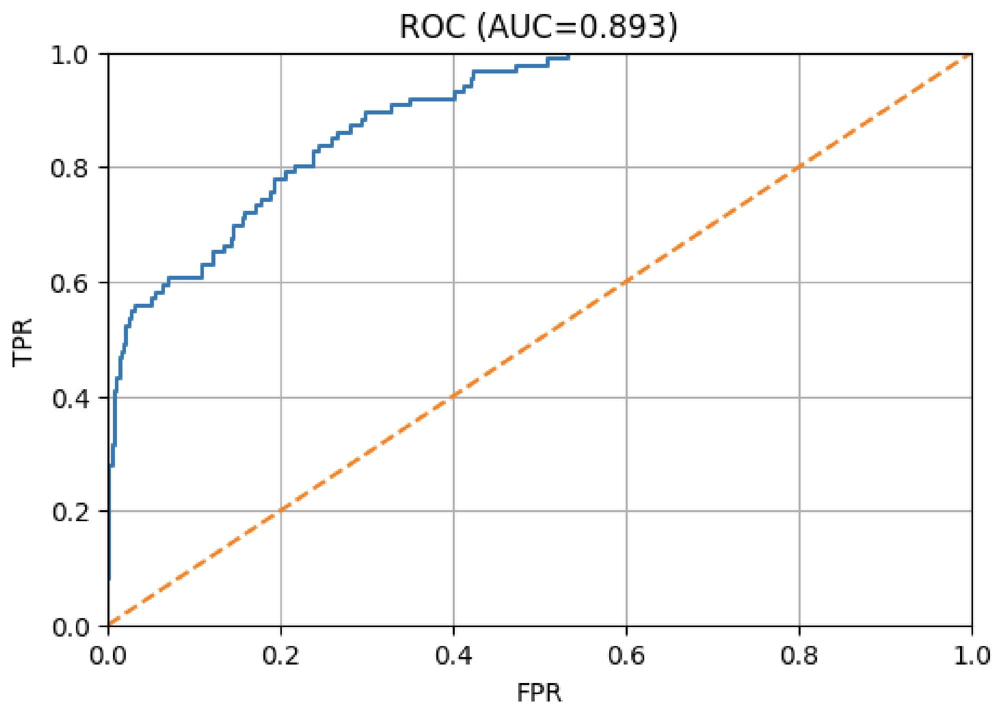
•The performance drops vs text-only. Common when:

- The text dominates and tabular adds noise.
- fusion method isn't optimal (I did weight 50/50).

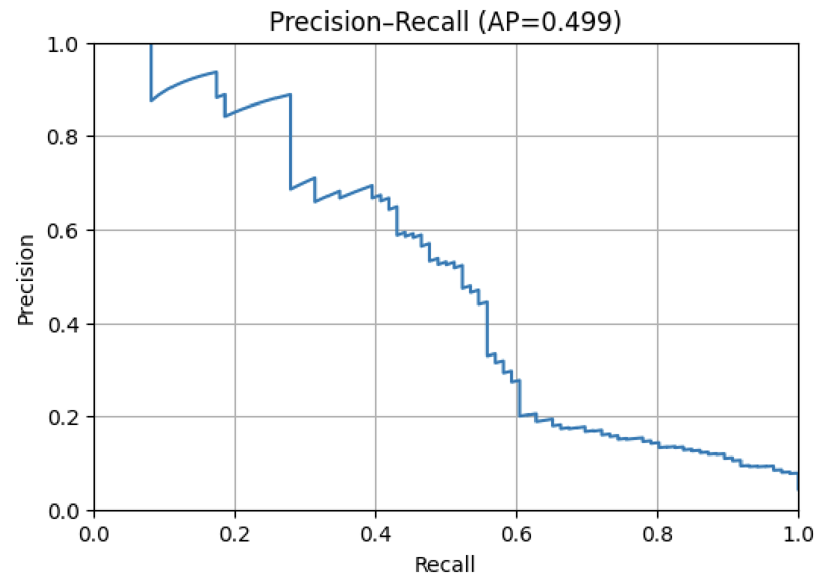
•Still: the fused model is very strong, PR-AUC ~0.52 is ~12× baseline

Budget-Constrained Evaluation

*Used test cost = 120 USD for this section



- **AUC=0.896** → strong separation between sick and healthy.
- At **10–20% false positives**, we capture **~70–80% of true cases**.
- This can be a **clinical sweet spot**, but the exact threshold should be chosen based on **risk tolerance and testing budget**.

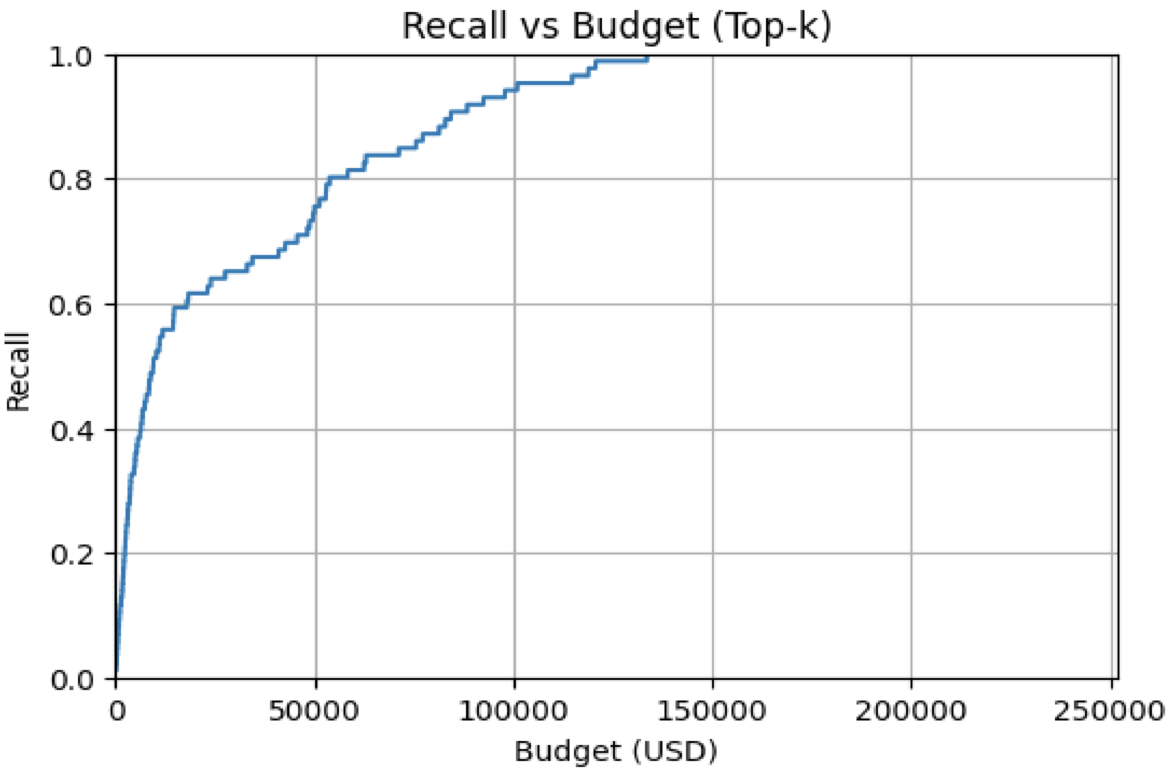


•In the test population, **baseline prevalence for hypertensive disorder is only ~4%.**

•**Random selection** of 100 women would detect ~4 true hypertensive cases.

•**Model-guided selection** of 100 women detects ~20 true cases — a **5× improvement in yield**.

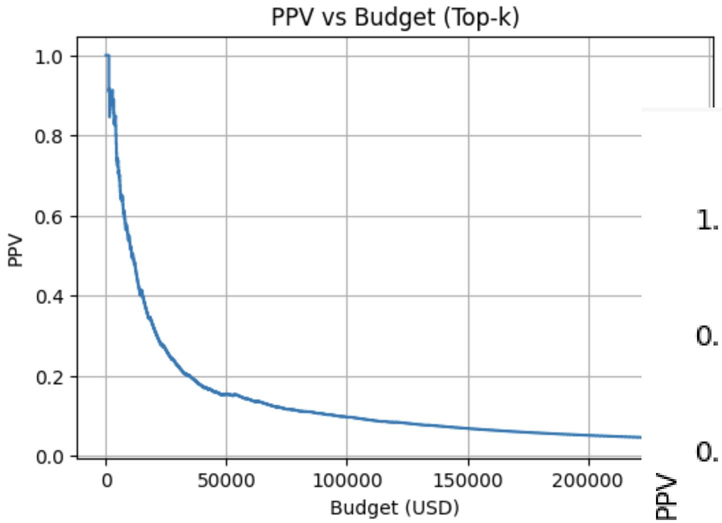
•This demonstrates that the model **concentrates scarce testing resources** on the women most likely to benefit, turning limited budget into much higher detection efficiency.

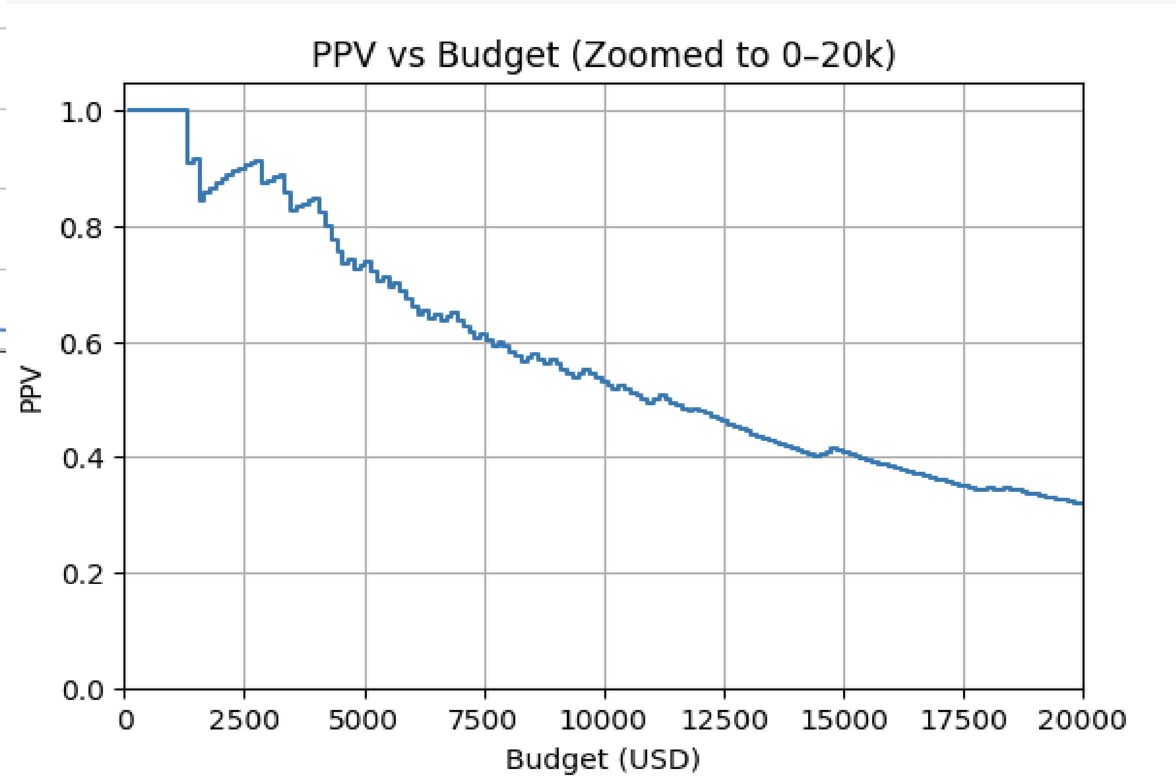


- As budget increases, recall rises quickly at first.
- After ~$100K, gains flatten → diminishing returns.

General conclusion:

- A **moderate budget (≈$50K)** already detects most cases (~75%).
- **Beyond $100K**, extra spending adds little benefit.





•**Too low budget (<$2k):** very high PPV, but you miss most true cases. Not practical clinically - With very low budget → PPV is high, but recall is low → only a few true cases are actually detected.

•**Moderate budget ($8k–12k):** PPV ~50–60%, still much higher than baseline 4%, and recall is already substantial.

•**Above $15k:** PPV drops below ~40%, so efficiency per test starts falling faster.

Next Steps:

•**Model benchmarking:** compare different models results.

•FE – check for more robust featue – change limits to get more sever high risk – for example age > 40 instead > 38

•**LLM for text features:** Extract features and set class risk factors (present / negated / not mentioned) using LLM's.

•**External validation:** test on independent cohort to ensure robustness.

This project shows how data science with advance AI tools can transform week-15 pregnancy data into a practical screening tool — one that not only predicts risk, but also helps allocate limited testing resources far more efficiently.