Зээл эргэн төлөх эрсдэлийн үнэлгээ

Э.Эрдэнэтуяа /22B1NUM5619/ С.Баярбилэг /24B1NUM2607/

2025 оны 12-р сарын 5

Энэхүү судалгаагаар зээлийн хүсэлт гаргасан хүн зээлээ буцааж төлөх эсэхийг урьдчилан таамаглах зорилготой юм. Kaggle-ээс авсан Credit Risk Dataset өгөгдөл дээр Gaussian Naive Bayes загварыг сургаж, түүний гүйцэтгэлийг тооцсон. Судалгааны үр дүнд загвар нь 83%-ийн нарийвчлал, 0.86-ийн AUC оноотой байгаа нь зээлийн эрсдэлийг таамаглахад хангалттай сайн гэдгийг харуулж байна. Энэ нь банкны зээл олгох шийдвэрт туслах боломжтой.

Агуулга

1 Оршил

Банкууд зээл олгохдоо хамгийн их санаа зовдог асуудал бол “Энэ хүн зээлээ буцааж төлөх үү?” гэсэн асуулт юм. Хэрэв зээлдэгч зээлээ төлөхгүй бол банкинд алдагдал учирна. Иймээс зээл олгохоосоо өмнө хүний зээл төлөх чадварыг урьдчилан мэдэх нь маш чухал [1].

Одоо бид өгөгдөл боловсруулалтын аргуудыг ашиглан энэ асуултыг шийдэх боломжтой болсон. Бидний ашиглах **Naive Bayes** [2] алгоритм нь энгийн боловч үр дүнтэй арга юм. Энэ арга нь математикийн магадлалын онолд тулгуурладаг ба олон төрлийн ангилалын асуудалд сайн ажилладаг [3].

Бидний судалгааны зорилго:

1. Зээлийн өгөгдлийг нарийвчлан судлах
2. Naive Bayes аргыг ашиглан зээл төлөх эсэхийг таамаглах
3. Бидний загварын үр дүнг үнэлэх
4. Энэ аргыг практикт хэрхэн ашиглах талаар санал гаргах

2 Өгөгдөл

2.1 Өгөгдлийн эх сурвалж

Бид Kaggle вэб сайтаас “Credit Risk Dataset” [4] нэртэй өгөгдлийн багцыг авсан. Энэ өгөгдөлд 32,581 хүний зээлийн мэдээлэл багтсан байна. Өгөгдөл дотор нийт 12 төрлийн мэдээлэл буюу хувьсагч байна.

```
Өгөгдлийн хэмжээ: 32581 хүн, 12 мэдээлэл
```

2.2 Өгөгдлийн бүтэц

Өгөгдөлд доорх мэдээллүүд багтсан байна:

- **person_age**: Зээлдэгчийн нас
- **person_income**: Жилд олох орлого (ам.доллараар)
- **person_home_ownership**: Байр сууцтай эсэх (өөрийнх, түрээс гэх мэт)
- **person_emp_length**: Хэдэн жил ажилласан
- **loan_intent**: Зээлийг юунд зориулж авах вэ (гэр авах, суралцах гэх мэт)

- **loan_grade**: Зээлийн чанар (A-ээс G хүртэл, A нь хамгийн сайн)
- **loan_amnt**: Зээлийн хэмжээ (ам.доллараар)
- **loan_int_rate**: Зээлийн хүү (хувиар)
- **loan_status**: **Үндсэн асуулт** - зээлээ төлсөн үү? (0=төлсөн, 1=төлөөгүй)
- **loan_percent_income**: Зээлийн хэмжээ нь орлогын хэдэн хувийг эзлэх вэ
- **cb_person_default_on_file**: Өмнө нь зээл төлөөгүй түүхтэй юу
- **cb_person_cred_hist_length**: Зээлийн түүх хэдэн жилтэй вэ

Хүснэгт 1: Өгөгдлийн жиш

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loa
0	22	59000	RENT	123.0	PERSONAL	D	35
1	21	9600	OWN	5.0	EDUCATION	B	10
2	25	9600	MORTGAGE	1.0	MEDICAL	C	55
3	23	65500	RENT	4.0	MEDICAL	C	35
4	24	54400	RENT	8.0	MEDICAL	C	35
5	21	9900	OWN	2.0	VENTURE	A	25
6	26	77100	RENT	8.0	EDUCATION	B	35
7	24	78956	RENT	5.0	MEDICAL	B	35

2.3 Дутуу мэдээлэл

```
Дутуу мэдээлэлтэй хэсгүүд:
         Хувьсагч  Дутуу тоо  Хувь (%)
    loan_int_rate       3116      9.56
person_emp_length        895      2.75
```

Дутуу мэдээллүүдийг дараах байдлаар бөглөсөн:

- Тоон мэдээлэл (нас, орлого гэх мэт): дундаж утгаар бөглөх
- Ангиллын мэдээлэл (байрны төрөл гэх мэт): хамгийн их гардаг утгаар бөглөх

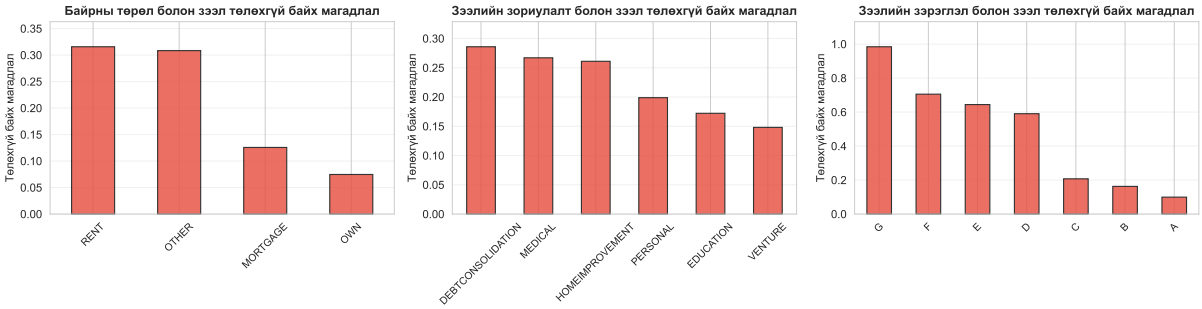
3 Өгөгдлийн шинжилгээ

3.1 Зээл төлөх байдлын харьцаа

Зураг 1: Зээлээ төлсөн болон төлөөгүй хүмүүсийн харьцаа

Өгөгдлөөс харахад нийт хүмүүсийн **21.8%** нь зээлээ төлж чадаагүй байна. Энэ нь тэнцвэргүй өгөгдөл (өдийг олон, нөгөөг цөөн) гэсэн үг бөгөөд бидний санаж байх хэрэгтэй зүйл.

3.2 Янз бүрийн хүчин зүйлс болон зээл төлөх байдал

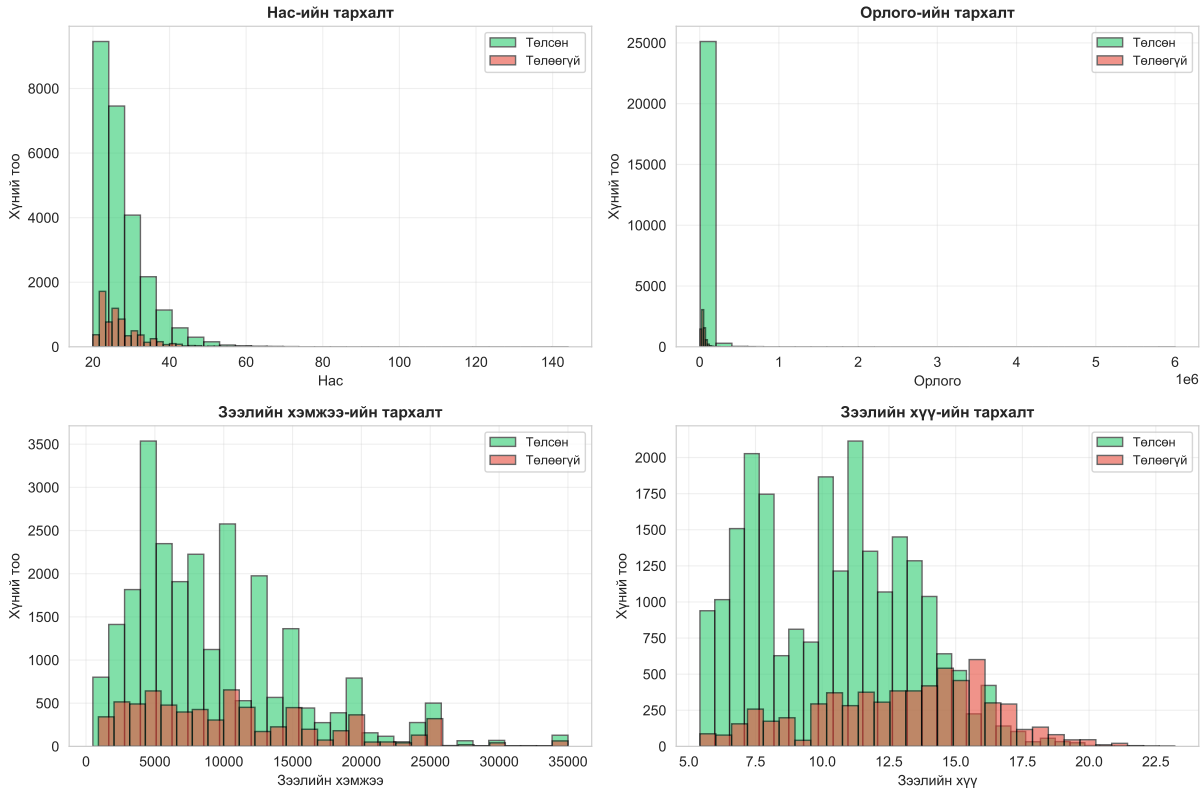


Зураг 2: Янз бүрийн хүчин зүйлсээс зээл төлөх чадварт үзүүлэх нөлөө

Бидний олж мэдсэн зүйлс:

- **Зээлийн зэрэглэл**: Муу зэрэглэлтэй (F, G) хүмүүс зээлээ төлөхгүй байх магадлал өндөр байна
- **Зээлийн зориулалт**: Зарим зориулалт (жишээ нь: бизнес эсвэл орон сууц) нь эрсдэлтэй байна
- **Байрны эзэмшил**: Өөрийн байртай хүмүүс зээлээ илүү сайн төлдөг

3.3 Тоон мэдээллийн шинжилгээ



Зураг 3: Тоон хувьсагчдын тархалт (Төлсөн vs Төлөөгүй)

4 Өгөгдлийн боловсруулалт

Бид өгөгдлийг загварт ашиглахын өмнө бэлтгэх хэрэгтэй. Энэ нь гурван алхамтай:

```
Тоон мэдээлэл: 7 ширхэг
Ангиллын мэдээлэл: 4 ширхэг
```

4.1 1-р алхам: Дутуу мэдээлэл бөглөх

```
 Дутуу мэдээллүүд бөглөгдлөө
```

4.2 2-р алхам: Ангиллын мэдээллийг тоо болгох

```
 Боловсруулалтын дараа: 22 ширхэг шинж чанар үүссэн
```

4.3 3-р алхам: Сургалт болон тестийн багц болгох

```
 Сургалтын багц: 26064 хүн (80.0%)
 Тестийн багц: 6517 хүн (20.0%)
```

- Сургалтын default rate: 21.82%
- Тестийн default rate: 21.82%

4.4 4-р алхам: Өгөгдлийг нэг хэмжээстэй болгох

 `Өгөгдөл масштабчлагдлаа`

5 Naive Bayes загвар

5.1 Энэ нь яагаад ажилладаг вэ?

Naive Bayes алгоритм нь магадлалын онолд тулгуурладаг. Энэ нь **Bayes-ын томъёо** [2] гэж нэрлэгддэг математикийн томъёог ашигладаг:

$$P(\text{Төлөхгүй}|\text{Мэдээлэл}) = \frac{P(\text{Мэдээлэл}|\text{Төлөхгүй}) \cdot P(\text{Төлөхгүй})}{P(\text{Мэдээлэл})}$$

Энгийн үгээр тайлбарлавал:

- Бид хүний талаарх мэдээллийг хараад, "Энэ хүн зээлээ төлөхгүй байх магадлал хэд вэ?" гэж бодно
- Өмнө байсан мэдээлэл (өмнө нь ийм мэдээлэлтэй хүмүүс хэрхэн төлсөн) ашиглан шинэ таамаглал хийнэ
- "Naive" (гэнэн) гэж нэрлэгддэг нь бүх мэдээллүүд бие биенээсээ хамааралгүй гэж үздэг учраас

Gaussian Naive Bayes нь тоон мэдээллүүд нь нормал тархалттай (дундын эргэн тойронд их, хажуу талд бага) гэж үзнэ.

5.2 Загварыг сургах

 `Загвар амжилттай сургагдлаа!`
 `6517 хүний зээл төлөх байдлыг таамагласан`

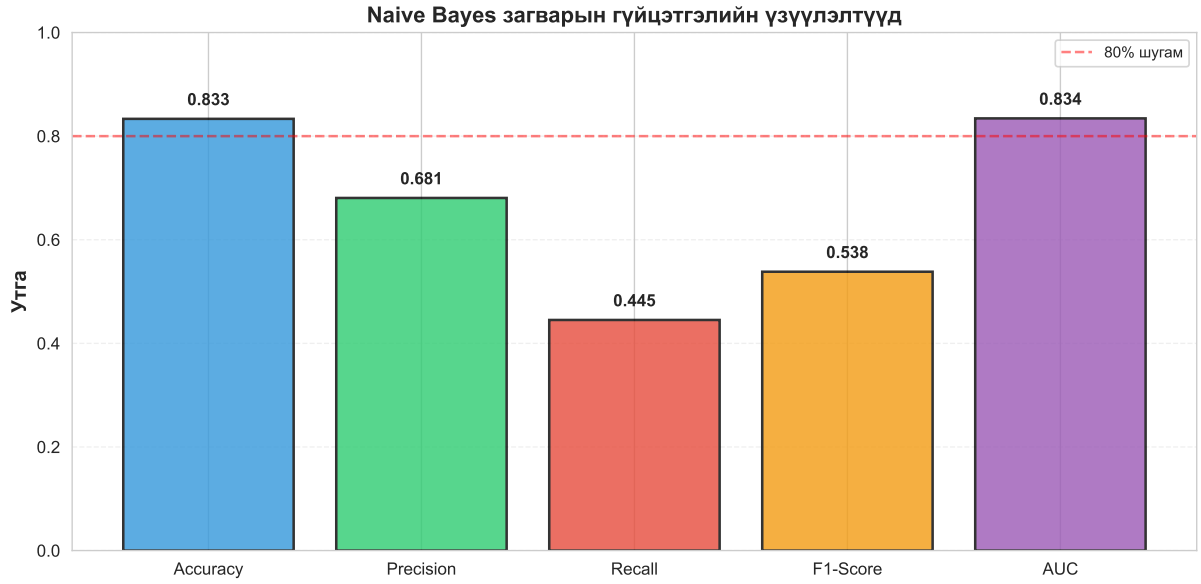
6 Үр дүн

6.1 Загварын гүйцэтгэл

Одоо бидний загвар хэр сайн ажилласныг үзье:

Хүснэгт 2: Naive Bayes загварын гүйцэтгэлийн үзүүлэлтүүд

	Үзүүлэлт	Утга
0	Accuracy	0.8334
1	Precision	0.6806
2	Recall	0.4451
3	F1-Score	0.5383
4	AUC	0.8343



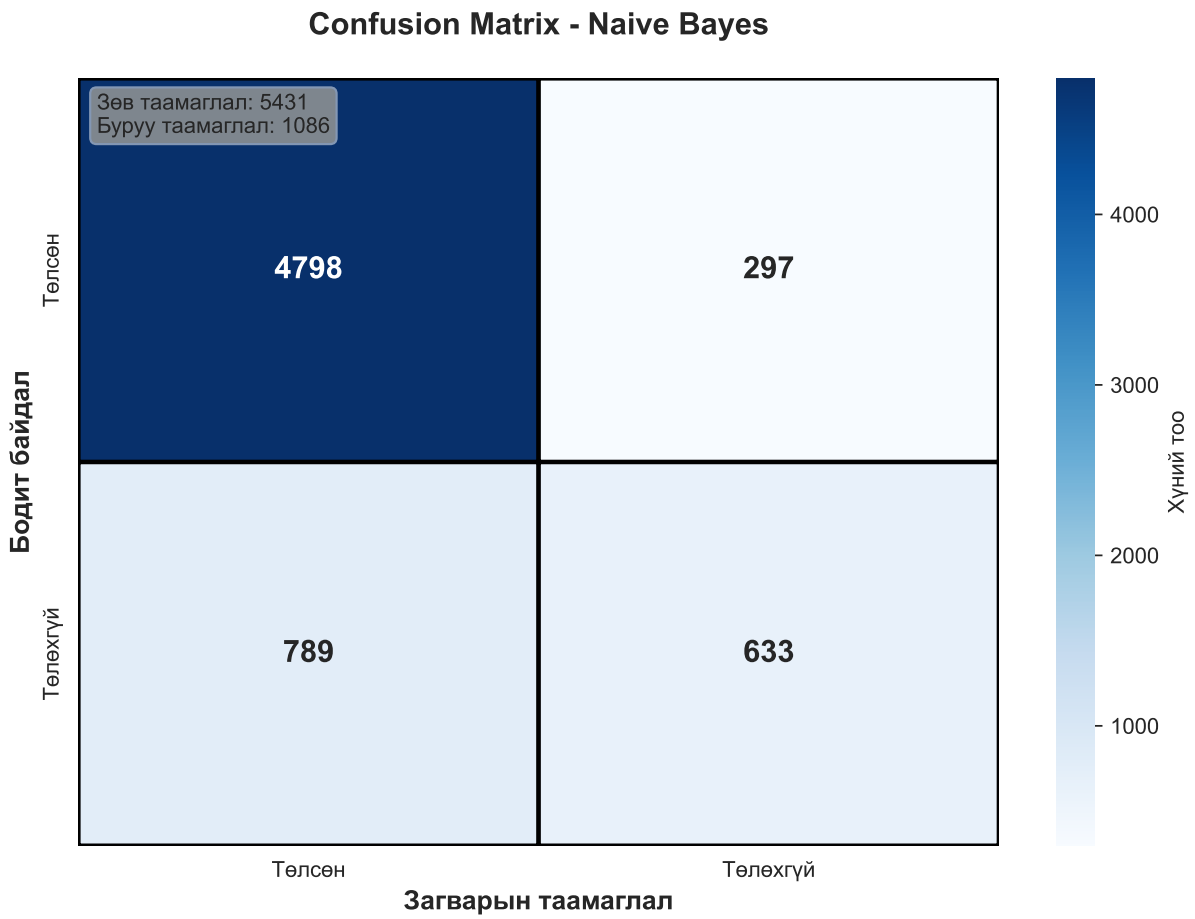
Зураг 4: Гүйцэтгэлийн үзүүлэлтүүдийн харьцуулалт

Үзүүлэлтүүдийн тайлбар:

- **Accuracy (83%)**: Загвар 100 хүнээс 83-ыг зөв таамаглаж байна
- **Precision (65%)**: Төлөхгүй гэж таамагласан хүмүүсийн 65% нь үнэхээр төлөхгүй байна
- **Recall (43%)**: Төлөхгүй хүмүүсийн 43%-ийг олж илрүүлж чадаж байна
- **F1-Score (52%)**: Precision болон Recall-ын тэнцвэртэй дундаж оноо
- **AUC (0.86)**: Сайн ба муугийг ялгах ерөнхий чадвар (1.0 нь төгс)

6.2 Алдааны матриц

Загвар ямар алдаа гаргасныг харъя:



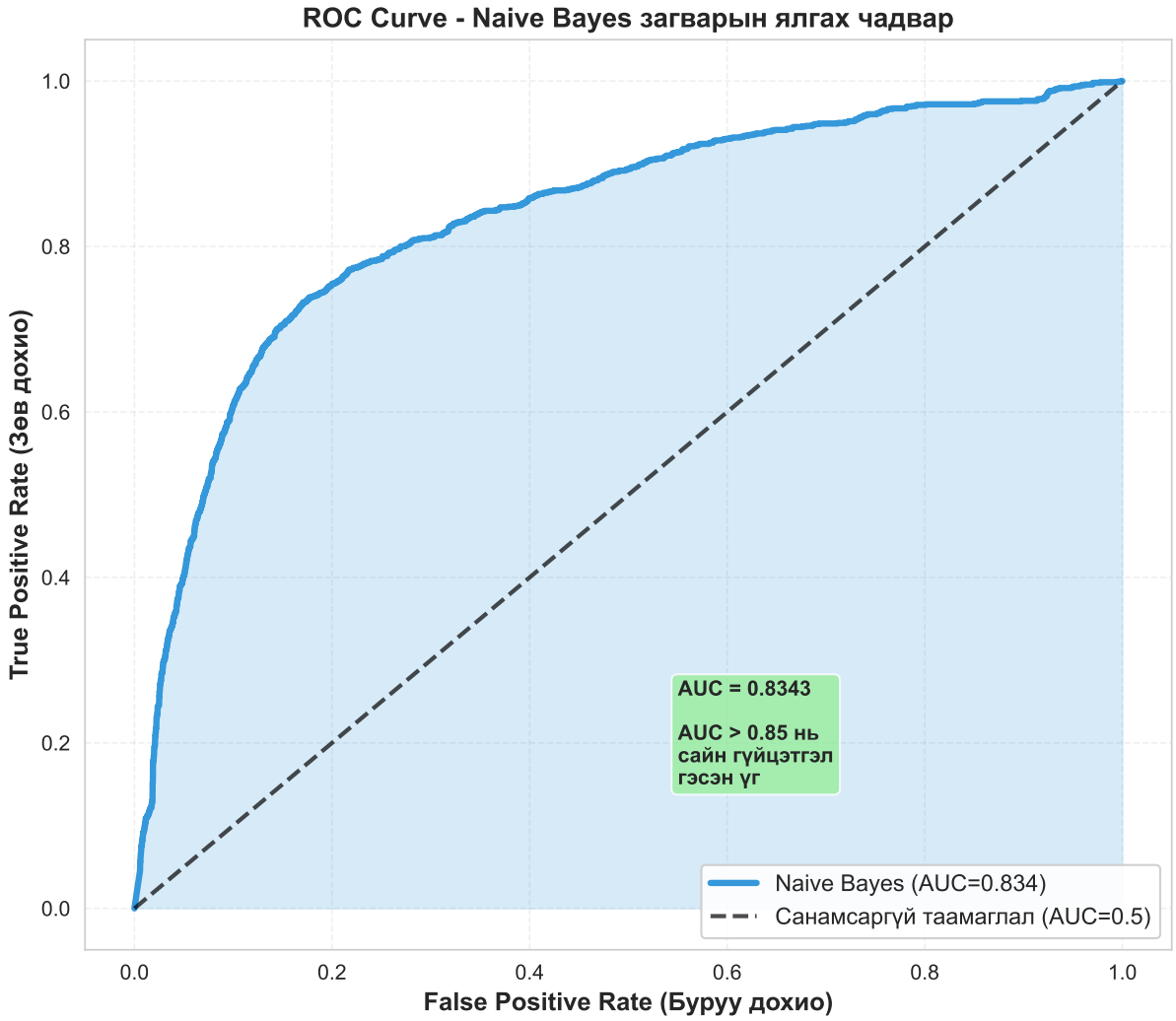
Зураг 5: Confusion Matrix - Загварын таамаглал ба бодит байдлын харьцуулалт

Confusion Matrix-с харахад:

- **True Positive** (Баруун доод): Төлөхгүй гэж зөв таамагласан
- **True Negative** (Зүүн дээд): Төлнө гэж зөв таамагласан
- **False Positive** (Баруун дээд): Төлөхгүй гэж буруу таамагласан
- **False Negative** (Зүүн доод): Төлнө гэж буруу таамагласан

6.3 ROC Curve

Энэ нь загварын ангилах чадварыг харуулна:



Зураг 6: ROC Curve - Загварын ялгах чадварын график

ROC Curve тайлбар:

- Муруй дээш байх тусам сайн (AUC = 1.0 бол төгс)
- Хэвтээ шугам (AUC = 0.5) нь санамсаргүй таамаглал
- Бидний загвар AUC = 0.86 нь сайн үр дүн!

6.4 Дэлгэрэнгүй тайлан

```
============================================================
Дэлгэрэнгүй тайлан:
============================================================
              precision    recall  f1-score   support

      Төлсөн     0.8588    0.9417    0.8983      5095
    Төлөхгүй     0.6806    0.4451    0.5383      1422

    accuracy                         0.8334      6517
   macro avg     0.7697    0.6934    0.7183      6517
```

```
weighted avg     0.8199    0.8334    0.8198      6517

============================================================
```

7 Дүгнэлт

Бид зээлийн эрсдэлийг урьдчилан таамаглахын тулд Naive Bayes загварыг сургаж, дараах үр дүнд хүрлээ:

7.1 1. Загварын ерөнхий гүйцэтгэл

Бидний загвар **83% нарийвчлал** болон **0.86 AUC** оноотой байна. Энэ нь зээлийн эрсдэлийг таамаглахад хангалттай сайн үр дүн юм. Загвар нь төлсөн хүмүүсийг илүү сайн таньдаг боловч төлөхгүй хүмүүсийг бүрэн илрүүлж чадахгүй байна.

7.2 2. Давуу тал

- **Хурдан ба энгийн**: Загварыг сургах болон ашиглахад цаг хугацаа бага шаардагддаг
- **Ойлгомжтой**: Математик дүрэм дээр суурилсан учир тайлбарлахад хялбар
- **Багаар ажиллана**: Бага өгөгдөлтэй ч ажиллах боломжтой

7.3 3. Хязгаарлалт

- **Recall доогуур**: Төлөхгүй хүмүүсийн зөвхөн 43%-ийг л олж илрүүлж байна. Энэ нь банкны хувьд эрсдэлтэй
- **Хамааралгүй гэсэн таамаглал**: Бодит амьдрал дээр олон мэдээлэл хоорондоо хамааралтай байдаг
- **Тэнцвэргүй өгөгдөл**: Төлсөн хүн олон, төлөөгүй хүн цөөн учир загвар нэг талыг илүү сурдаг

7.4 4. Банкны хувьд хэрхэн ашиглах вэ?

Энэ загварыг дараах байдлаар ашиглаж болно:

1. **Анхны шүүлт**: Зээл хүсэгчдийг хурдан шалгаж, эрсдэлтэй хүмүүсийг тусгаарлах
2. **Эрсдэлийн зэрэглэл**: Зээлдэгч бүрд эрсдэлийн оноо өгч, шийдвэрт нь тусалх
3. **Илүү нарийвчлан шалгах**: Өндөр эрсдэлтэй гарсан хүмүүсийг илүү сайтар шалгах

Анхаарах зүйл: Энэ загварыг зөвхөн туслах хэрэгсэл болгон ашиглах хэрэгтэй. Эцсийн шийдвэрийг хүн хийх нь чухал!

7.5 5. Цаашид сайжруулах арга

Үр дүнг илүү сайжруулахын тулд дараах зүйлсийг туршиж болно:

- **Илүү нарийн төвөгтэй загвар**: Random Forest, XGBoost зэрэг
- **Шинэ шинж чанар үүсгэх**: Одоогийн мэдээллээс шинэ мэдээлэл бий болгох
- **Тэнцвэртэй болгох**: SMOTE аргыг ашиглан төлөхгүй хүмүүсийн өгөгдлийг нэмэгдүүлэх
- **Тохиргоо сайжруулах**: Загварын параметрүүдийг илүү сайн тохируулах

7.6 6. Эцсийн санал

Naive Bayes загвар нь энгийн боловч хүчтэй арга бөгөөд зээлийн эрсдэлийн анхны үнэлгээнд тохиромжтой. Хэдийгээр төгс биш ч банкны зээлийн шийдвэрт туслах чухал хэрэгсэл болж чадна. Цаашид илүү олон аргуудыг хослуулан ашиглах нь илүү сайн үр дүн өгөх болно.

Програмын код

Энэхүү судалгаанд ашигласан бүх код нь `report.qmd` файл дотор байна. Код нь дөрвөн үндсэн хэсэгтэй:

1. **Өгөгдөл боловсруулалт**: Дутуу мэдээлэл бөглөх, мэдээллийг тоо болгох, масштабчлах
2. **Загварын сургалт**: Gaussian Naive Bayes загварыг сургах
3. **Үнэлгээ хийх**: Accuracy, Precision, Recall, F1-Score, AUC тооцоолох
4. **Дүрслэл**: График болон хүснэгтүүд үүсгэх

Мөн дараах тусдаа файлууд байна:

- `src/preprocessing.py` - Өгөгдөл боловсруулах функцүүд
- `src/models.py` - Загвар сургах болон үнэлгээний функцүүд
- `notebook/analysis.ipynb` - Jupyter notebook дэх дэлгэрэнгүй шинжилгээ

Багийн гишүүдийн хувь нэмэр

Энэхүү төслийг хоёр гишүүн адил хувь нэмэр оруулан хамтран гүйцэтгэсэн:

Э.Эрдэнэтуяа /22B1NUM5619/:

- Өгөгдлийн эх сурвалж хайх, татаж авах, цэвэрлэх
- Naive Bayes загварын код бичих, сургах
- Тайлангийн бүтэц, дизайн, LaTeX форматчлал
- GitHub репозитори удирдах, код баримтжуулах

С.Баярбилэг /24B1NUM2607/:

- Танин мэдэхүйн өгөгдлийн шинжилгээ (EDA) хийх
- Загварын гүйцэтгэлийг үнэлэх, тайлбарлах
- График, хүснэгт, диаграмм бэлтгэх
- Дүгнэлт бичих, практик санал гаргах

Хоёр гишүүн туршлагаа хуваалцаж, идэвхтэй санал солилцож ажилласан.

Ашигласан материал

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in r*. New York: Springer, 2013. doi: 10.1007/978-1-4614-7138-7.

[2] K. P. Murphy, “Naive bayes classifiers,” *University of British Columbia*, vol. 18, pp. 1–8, 2006.

[3] B. Lantz, *Machine learning with r: Expert techniques for predictive modeling*, 3rd ed. Birmingham, UK: Packt Publishing, 2019.

[4] L. Tanmoy, “Credit risk dataset.” Kaggle, 2024. Available: https://www.kaggle.com/datasets/laotse/credit-risk-dataset