Big Data Analysis Competition Report

- Credit Card Approval Analysis-

Advisor Professor: 배재권

-MangoMango Team-

사학과 - 조진목

경영정보학전공 - 이정호

경영정보학전공 - NGUYEN THANH TUNG

보고서 요약

금융 분야의 역동적이고 경쟁적인 환경에서 은행에게 신용카드 승인 과정을 최적화하는 것은 매우 중요합니다. 이 보고서는 신용카드 후보자들의 신청 데이터를 조사하여 승인율에 영향을 미치는 요인들을 파악합니다. 우리는 전 세계 다양한 지역의 신청자들로부터의 포괄적인 재정적 및 개인적 정보를 포함하는 25,128건의 데이터셋을 분석했습니다.

고급 분석 및 기계 학습 기술을 사용하여 신청자의 신용도를 정확하게 평가하는 강력한 예측 모델을 개발했습니다. 이 모델은 재정 안정성, 신용 기록 및 기타 관련 요인들의 상세한 분석을 사용하여 신청자가 신용카드 지불을 기본으로 할 가능성을 효과적으로 식별합니다. 이를 통해 은행은 보다 정보에 입각한 정밀한 신용카드 승인 결정을 내릴 수 있습니다.

이 연구의 발견은 신용 기본 위험을 줄이는 것뿐만 아니라 승인 과정을 간소화하는 데에도 기여합니다. 이는 은행의 운영 효율성과 고객 만족도 수준을 크게 향상시킵니다. 이 연구에서 얻은 통찰력은 신중한 신용 위험 관리 관행을 형성하고 은행 부문의 지속 가능한 성장을 촉진하는 데 중요한 역할을 합니다.

소개

은행 업계의 동적이고 변화하는 세계에서 신용카드 승인 프로세스를 최적화하는 것은 성장과 경쟁력을 유지하기 위해 중요합니다. 연구에 따르면, 신용카드 승인의 효율적인 관리는 은행의 재무 성과와 고객 만족도를 크게 향상시킬 수 있습니다. 잘못된 신용 승인의 비용 영향은 종종 고객 유지 전략과 관련된 비용을 훨씬 능가합니다.

우리는 실제 상황에서의 관련성과 데이터의 포괄적인 특성으로 인해 25,128개의 레코드로 구성된 데이터 집합을 분석하기로 결정했습니다. 이 데이터 집합은 다양한 지원자 풀을 반영뿐만 아니라 금융 역사부터 개인 정보까지 각 지원자에 대한 심층 정보를 제공합니다. 이러한 풍부한 데이터는 신용카드 승인 결정에 영향을 미치는 역학을 깊이 이해하게 해주며 은행이 더 정확하고 고객 중심의 신용 결정을 내릴 수 있게 합니다.

이 보고서에서 우리의 접근 방식은 금융 데이터를 활용한 강력한 기계 학습 모델을 구현하는 것이며, 주목할만한 신용카드 승인 예측 능력에 중점을 둡니다. 목표는 정확성뿐만 아니라 실제 은행 과제에 대한 실용적인 적용 가능성에서도 뛰어난 모델을 개발하는 것입니다.

보고서는 탐색적 데이터 분석부터 데이터 처리, 피처 엔지니어링 및 예측 모델 개발 방법론에 이르기까지 수행한 프로세스에 대한 세부 정보를 제공합니다. 또한 고객 이탈을 가장 잘 예측하는 재무 지표를 강조하는 결과를 제시할 것입니다. 이 분석을 통해 은행의 고객 유지 역학에 대한 가치 있는 통찰력을 제공하고 이탈을 완화하기 위한 실행 가능한 전략을 제안하는 것이 목표입니다.

데이터 수집 및 변수 설명

1.데이터 수집

이 보고서에서, 우리는 Kaggle에 공개된 신용 카드 승인 예측에 관한 데이터 세트를 사용합니다. 이 데이터 세트는 신용 카드 승인 결정에 영향을 미치는 다양한 요소들, 금융 이력부터 사용자의 개인 정보에 이르기까지, 자세하고 다양한 정보를 포함하고 있습니다. 이 데이터 세트에 접근하고 분석하기 위해, 우리는 데이터를 세심하고 체계적으로 전처리하고 통합하는 작업을 수행했습니다. 데이터 통합 후에는 총 25,128개의 레코드와 21개의 변수를 포함하게 되었습니다.

2.변수 설명

Applicant_ID: 신청자의 고유 식별 번호
Applicant_Gender: 신청자의 성별.
Owned_Car: 신청자가 자동차를 소유하고 있는지 여부를 나타냅니다.
Owned_Realty: 신청자가 재산을 소유하고 있는지 여부를 나타냅니다.
Total_Children: 신청자의 자녀 수입니다.
Total_Income: 신청자의 연간 총 수입입니다.
Income_Type: 신청자의 소득 유형.
Education_Type: 신청자의 교육 수준.
Family_Status: 신청자의 결혼 상태나 가족 상황.
Housing_Type: 신청자의 주거 형태.
Owned_Mobile_Phone: 신청자가 휴대전화를 소유하고 있는지 여부.
Owned_Work_Phone: 신청자가 업무용 전화를 소유하고 있는지 여부.
Owned_Phone: 신청자가 개인 전화(휴대전화 외)를 소유하고 있는지 여부.
Owned_Email: 신청자가 이메일 주소를 소유하고 있는지 여부.
Job_Title: 신청자의 직업 타이틀.
Total_Family_Members: 신청자 가족의 총 인원 수.
Applicant_Age: 신청자의 나이.
Years_of_Working: 신청자의 총 근무 연수.
Total_Bad_Debt: 신청자의 총 나쁜 빚의 양.
Total_Good_Debt: 신청자의 총 좋은 빚의 양.
Status: 신용 카드 승인 신청의 최종 상태(1 : 승인, 0 : 거절).

탐색적 데이터 분석

```
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Applicant_ID          25128 non-null  int64
 1   Applicant_Gender      25128 non-null  object
 2   Owned_Car             25128 non-null  int64
 3   Owned_Realty          25128 non-null  int64
 4   Total_Children        25128 non-null  int64
 5   Total_Income          25128 non-null  int64
 6   Income_Type           25128 non-null  object
 7   Education_Type        25128 non-null  object
 8   Family_Status         25128 non-null  object
 9   Housing_Type          25128 non-null  object
 10  Owned_Mobile_Phone    25128 non-null  int64
 11  Owned_Work_Phone      25128 non-null  int64
 12  Owned_Phone           25128 non-null  int64
 13  Owned_Email           25128 non-null  int64
 14  Job_Title             25128 non-null  object
 15  Total_Family_Members  25128 non-null  int64
 16  Applicant_Age         25128 non-null  int64
 17  Years_of_Working      25128 non-null  int64
 18  Total_Bad_Debt        25128 non-null  int64
 19  Total_Good_Debt       25128 non-null  int64
 20  Status                25128 non-null  int64
```

제공된 데이터에 따르면 데이터 세트에는 총 21개의 변수가 포함되어 있습니다. 'Applicant_Gender', 'Income_Type', 'Education_Type', 'Family_Status', 'Housing_Type', 그리고 'Job_Title' 등 특정 6개 변수는 'object' 유형입니다. 나머지 모든 변수는 'int64' 데이터 유형입니다. 데이터 세트 검사 결과, 25,128개의 기록이 있으며 누락되거나 중복된 값이 없는 것으로 나타나, 후속 분석 단계에 대해 데이터가 준비되어 있음을 시사합니다.

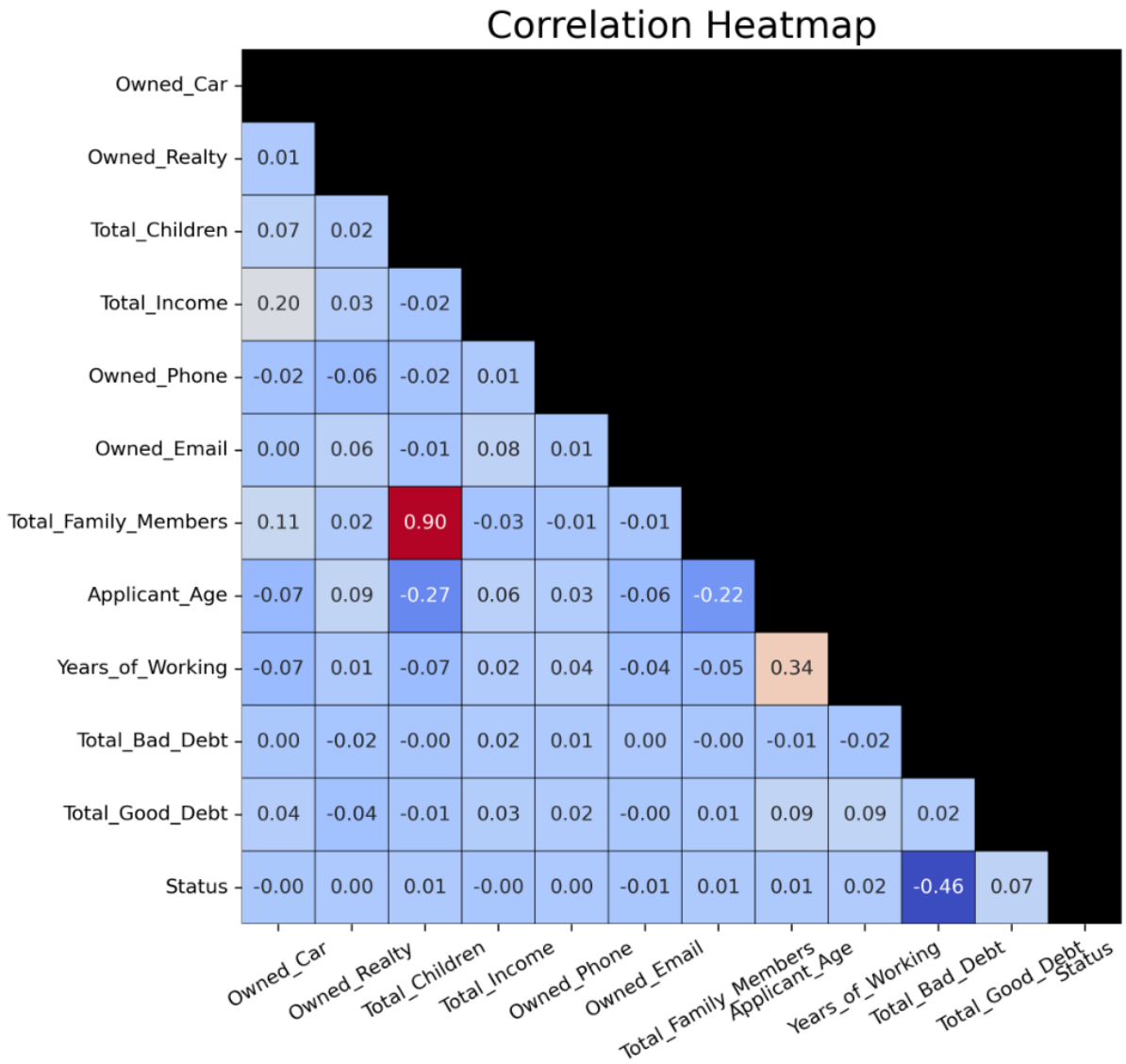
1.기술 통계

	count	mean	std	min	25%	50%	75%	max
Applicant_ID	25128.00	5078835.48	41943.78	5008806.00	5042225.75	5079004.00	5115603.25	5150487.00
Owned_Car	25128.00	0.42	0.49	0.00	0.00	0.00	1.00	1.00
Owned_Realty	25128.00	0.65	0.48	0.00	0.00	1.00	1.00	1.00
Total_Children	25128.00	0.51	0.76	0.00	0.00	0.00	1.00	5.00
Total_Income	25128.00	194836.50	104521.12	27000.00	135000.00	180000.00	225000.00	1575000.00
Owned_Mobile_Phone	25128.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Owned_Work_Phone	25128.00	0.27	0.45	0.00	0.00	0.00	1.00	1.00
Owned_Phone	25128.00	0.29	0.46	0.00	0.00	0.00	1.00	1.00
Owned_Email	25128.00	0.10	0.30	0.00	0.00	0.00	0.00	1.00
Total_Family_Members	25128.00	2.29	0.93	1.00	2.00	2.00	3.00	7.00
Applicant_Age	25128.00	41.00	9.55	21.00	33.00	40.00	48.00	68.00
Years_of_Working	25128.00	7.69	6.42	1.00	3.00	6.00	10.00	44.00
Total_Bad_Debt	25128.00	0.33	1.57	0.00	0.00	0.00	0.00	49.00
Total_Good_Debt	25128.00	21.06	14.74	1.00	9.00	18.00	31.00	61.00
Status	25128.00	1.00	0.07	0.00	1.00	1.00	1.00	1.00

신용 카드 승인 데이터 세트에 대한 기술적 분석을 통해 핵심 변수들에 대한 중요한 통계적 특성을 확인하였습니다. 우선, 신청자의 총 수입(Total_Income)은 평균 약 194,836으로, 상당한 편차를 보이는(표준편차 104,521) 주요 지표입니다. 이는 신청자 간의 경제적 격차가 크다는 것을 시사하며, 신용 승인 과정에서 중요한 고려 사항으로 작용할 수 있습니다. 또한, 신청자의 평균 나이(Applicant_Age)는 41세로, 신용 카드 발급 대상의 연령 분포를 나타내며, 평균 근무 연수(Years_of_Working)는 약 7.69년으로, 신청자의 직업 안정성과 경력을 반영하는 지표로 해석될 수 있습니다.

신용 카드 승인 상태(Status)는 본 데이터 세트에서 평균 1.00을 기록하여, 대부분의 신청자가 신용 카드 승인을 받았음을 나타냅니다. 이는 선택된 표본이 신용 카드 발급에 긍정적인 결과를 보인 고객군을 대표할 수 있음을 시사합니다. 이와 더불어, 신청자가 자동차(Owned_Car) 및 부동산(Owned_Realty) 소유 여부는 각각 평균 0.42와 0.65로, 신청자의 자산 상태와 신용도에 영향을 미칠 수 있는 중요한 요소로 판단됩니다.

2상관 계수



Total_Family_Members와 Applicant_Age: 가족 구성원 수와 신청자의 나이 사이에는 강한 양의 상관관계(r = 0.90)가 있으며, 이는 나이가 많은 신청자일수록 가족 구성원 수가 더 많을 수 있음을 나타냅니다.

Applicant_Age와 Total_Good_Debt: 신청자의 나이와 총 좋은 빚 사이에는 중간 정도의 음의 상관관계(r = -0.27)가 있어, 신청자의 나이가 많아질수록 '좋은 빚'이 줄어들 수 있음을 시사합니다. 이는 나이가 많은 신청자들이 신용을 덜 활용하거나 빚을 더 많이 상환했음을 의미할 수 있습니다.

Applicant_Age와 Years_of_Working: 신청자의 나이와 근무 연수 사이에도 중간 정도의 음의 상관관계(r = -0.22)가 있어, 나이가 많은 신청자들은 은퇴에 가까워 근무 연수가 데이터 세트에 적게 기록될 수 있음을 암시

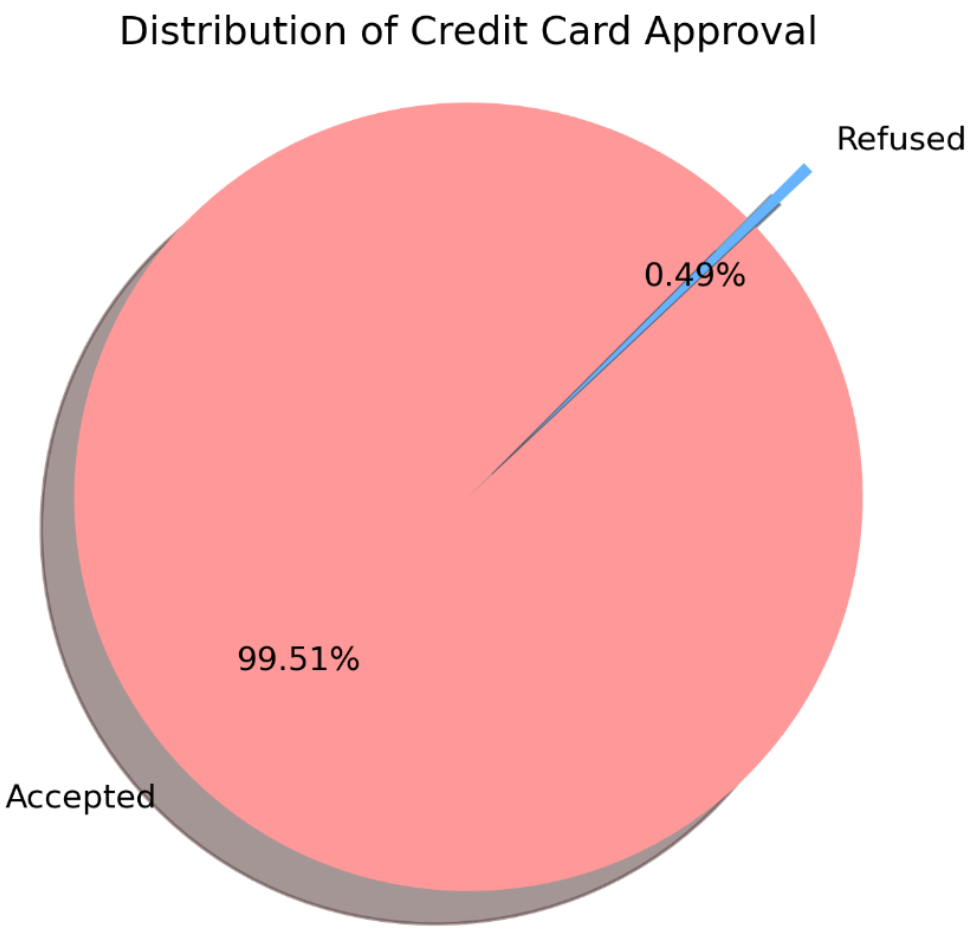
합니다.

Years_of_Working와 Total_Good_Debt: 근무 연수와 총 좋은 빚 사이에는 양의 상관관계(r = 0.34)가 있어, 근무 연수가 긴 사람들은 더 많은 좋은 빚을 가지고 있으며, 이는 더 긴 고용 기록으로 인한 신용도를 반영할 수 있습니다.

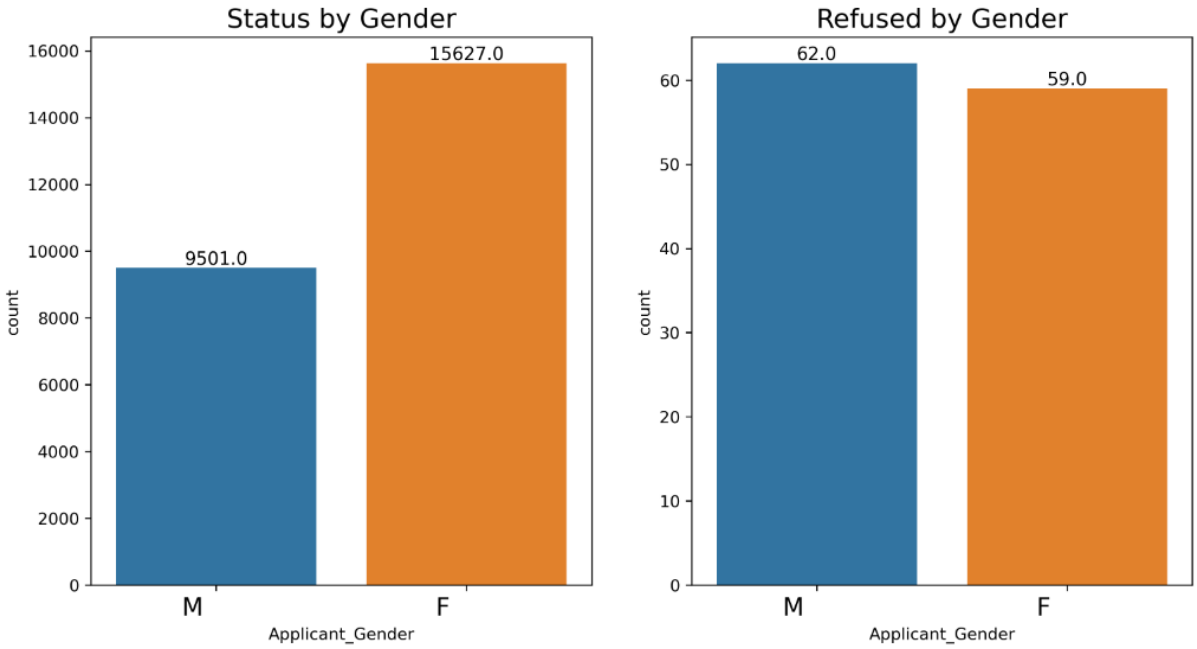
Status와 Total_Good_Debt: 신용 카드 승인 상태(승인을 나타내는 상태가 1임)와 총 좋은 빚 사이의 음의 상관관계(r = -0.46)는 이 데이터 세트에서 좋은 빚이 많을수록 신용 카드 승인 가능성이 낮음을 의미합니다.

기타 변수들 간의 상관관계: 나머지 변수들과 신용 카드 승인 상태와의 상관관계는 매우 낮으며, 계수가 거의 0에 가깝습니다. 이는 관찰된 데이터 세트 내에서 신용 카드 승인 상태와 선형 관계가 매우 약하거나 없음을 나타냅니다.

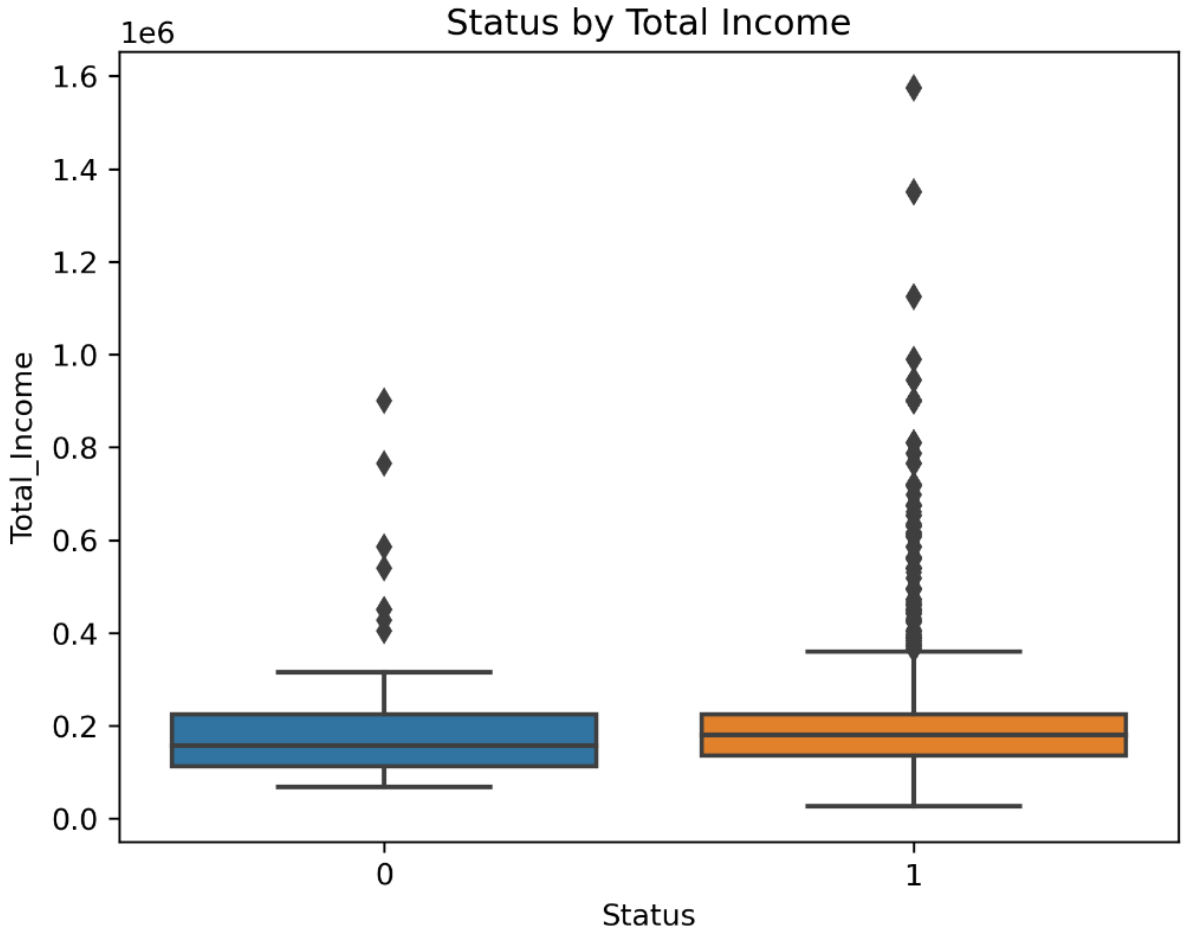
3. 변수 분석



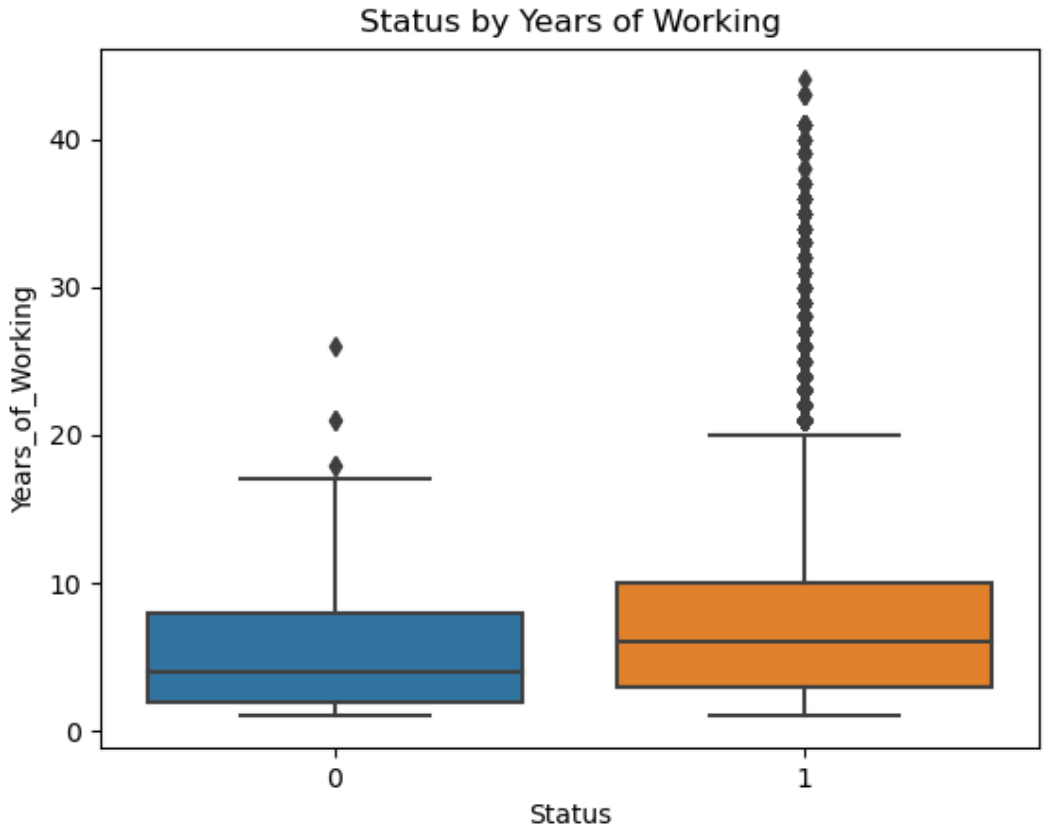
"Credit Card Approval" 분포도 분석에서 신용카드 승인 데이터에 상당한 불균형이 드러났습니다. 비정상적으로 높은 99.51%의 승인 비율에 비해, 거절된 신청은 단지 0.49%에 불과합니다. 이러한 불균형은 효과적인 예측 모델을 구축하는 데 어려움을 초래할 수 있으며, 모델이 현실에서 정확하게 식별해야 하는 거절 사례를 감지하는 데 민감하지 않게 만들 수 있습니다. 이 문제를 해결하기 위해, 우리는 filtering 방법으로 데이터 전처리를 진행한 다음, 승인 사례가 적은 경우에는 오버샘플링을, 승인 사례가 많은 경우에는 언더샘플링을 적용하는 등의 데이터 재구성 기술을 사용하여 데이터 균형을 개선하고 모델의 예측 능력을 강화하고자 합니다.



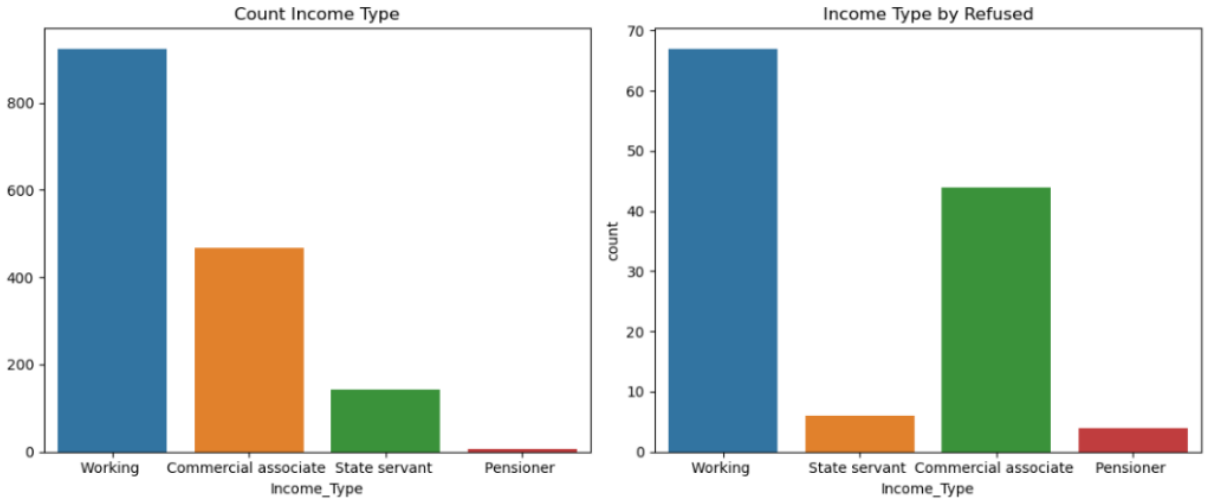
"성별에 따른 신용카드 승인 현황" 차트는 신용카드 신청서가 성별에 따라 어떻게 분포되는지를 보여줍니다. 구체적으로 여성(F)의 신청서 수가 15,627건으로 남성(M)의 9,501건보다 더 많습니다. 그러나 "성별에 따른 거절 현황" 차트는 두 성별 간에 거절된 신청서 수가 거의 균형을 이루고 있음을 보여줍니다. 남성은 62건, 여성은 59건이 거절되었습니다. 이는 남성이 여성보다 거절될 가능성이 더 높음을 시사합니다



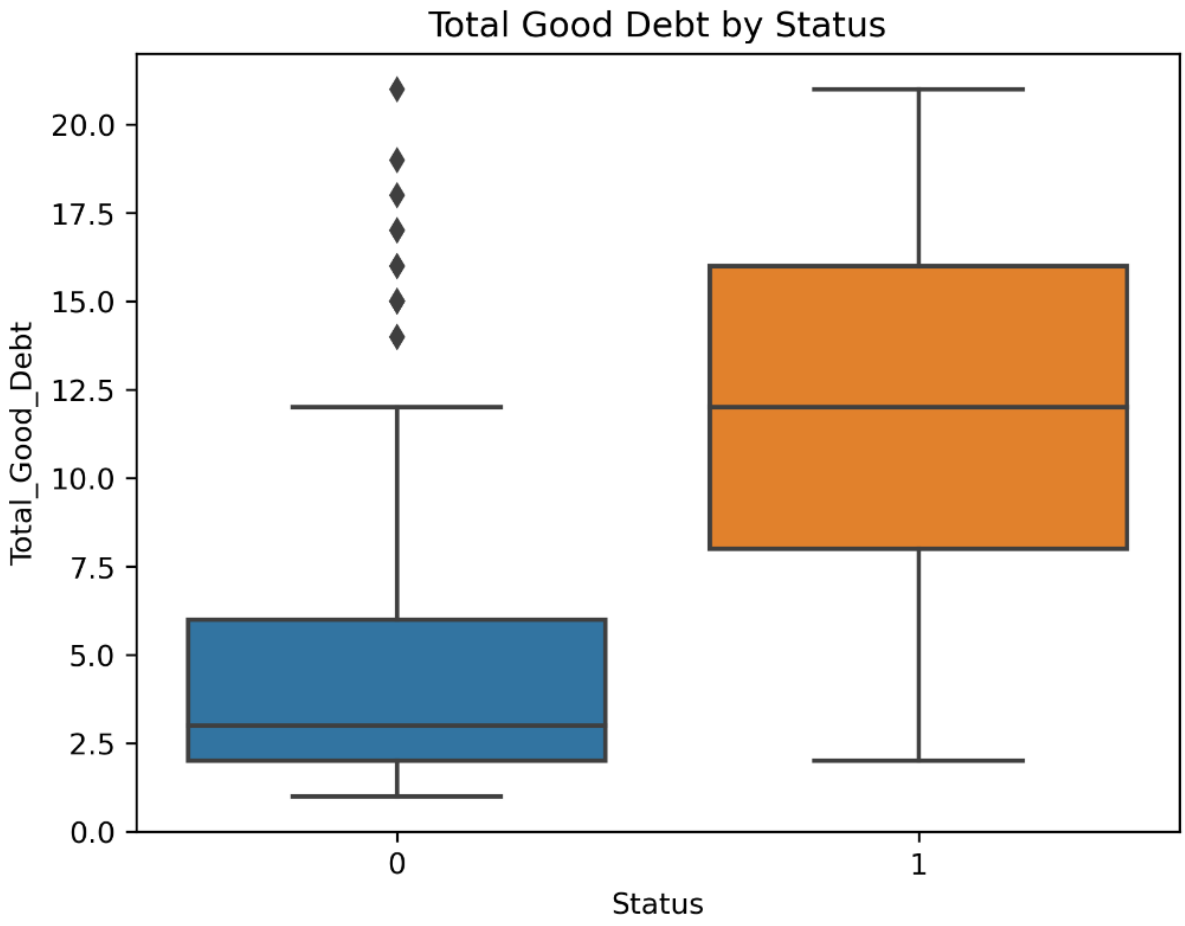
"Status by Total Income" 차트는 신용카드 승인 결과에 따른 고객의 총 소득 분포를 명확하게 보여줍니다. 승인된 그룹은 더 넓은 소득 범위를 가지고 있으며, 평균 소득이 더 높고 "매우 높은" 소득을 가진 고객들은 거의 다 승인되었습니다.

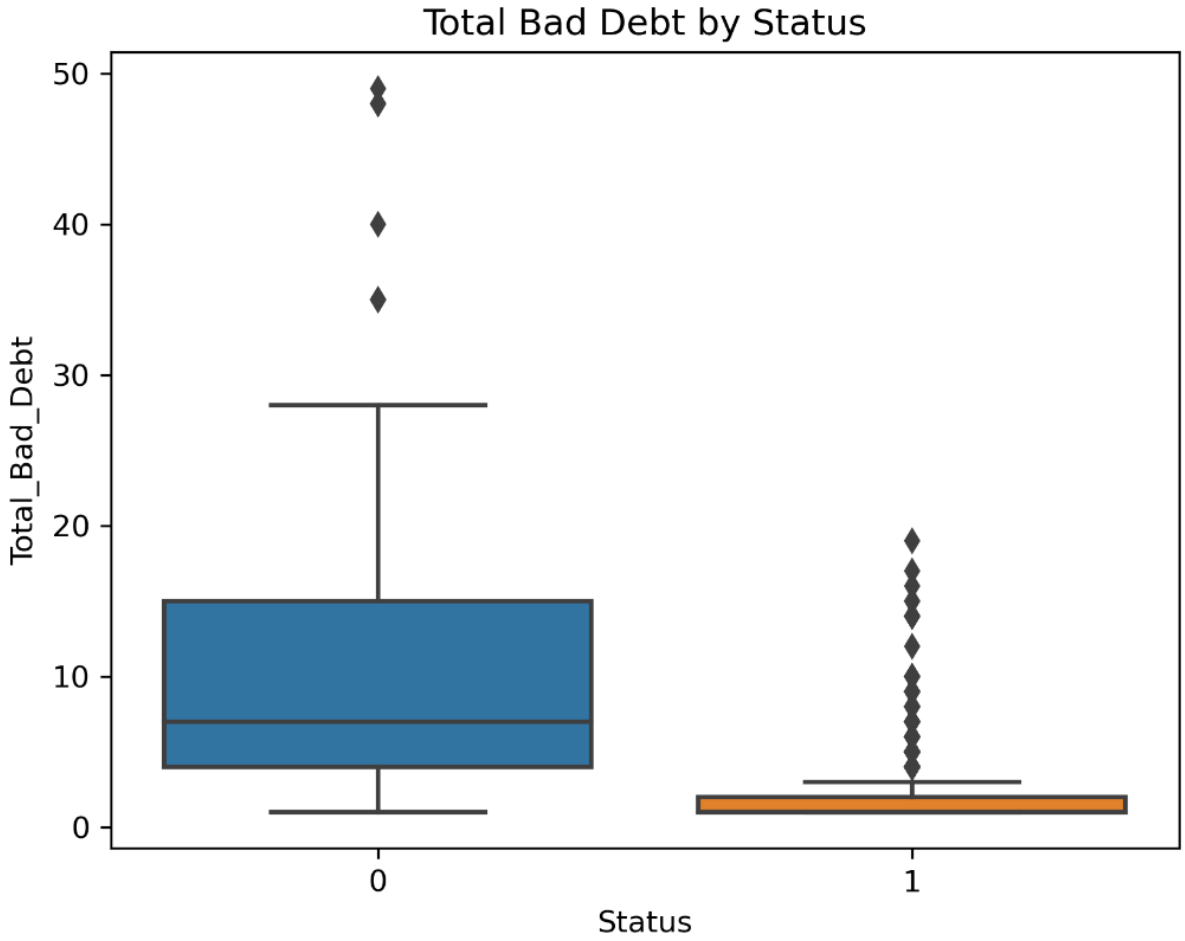


"Status by Years of Working" 차트는 업무 경력의 연수가 신용카드 승인 결정에 미치는 영향을 보여줍니다. 데이터에 따르면 장기간 근무한 사람들은 일반적으로 승인 확률이 높다는 것을 나타냅니다. 이는 승인된 그룹의 중앙값이 더 높고 승인된 그룹의 분산 범위가 더 넓다는 것으로 확인됩니다. 또한 차트는 28세 이상인 사람들은 모두 승인을 받는다는 것을 보여줍니다.



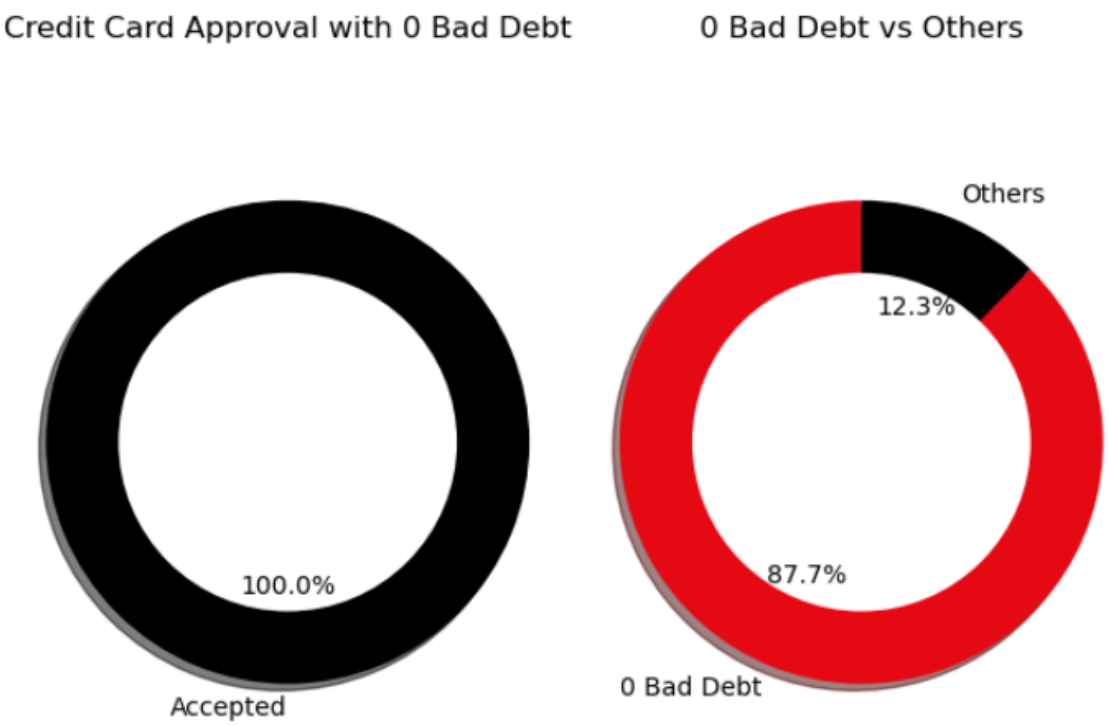
본 "Count Income Type" 및 "Income Type by Refused" 차트를 통해 Income Type의 분포를 볼 수 있습니다. 여기서 큰 차이점은 "Commercial associate"가 더 자주 거절되는 경향이 있고, "State servant"는 덜 거절되는 경향이 있다는 것입니다.





두 차트 "Total Good Debt by Status"와 "Total Bad Debt by Status"를 통해 승인과 거절된 두 그룹 사이에 뚜렷한 차이를 볼 수 있습니다. "Total Good Debt by Status" 차트에서는 승인된 고객들이 거절된 그룹에 비해 평균 좋은 빚이 더 많으며, 중요한 것은 그들의 좋은 빚의 총액이 항상 0보다 큽니다. 이는 좋은 빚이 대출 승인 결정에 긍정적인 요소일 수 있음을 나타냅니다.

반면에, "Total Bad Debt by Status" 차트는 대부분의 승인된 고객들이 낮은 나쁜 빚의 총액을 가지고 있으며, 통상 3 이하입니다. 반면에 거절된 고객들은 훨씬 더 높은 나쁜 빚을 가지고 있습니다. 이는 나쁜 빚이 대출 승인 가능성에 큰 영향을 미치는 요소일 수 있으며, 나쁜 빚이 많은 사람들은 거절될 가능성이 더 높음을 나타낼 수 있습니다.

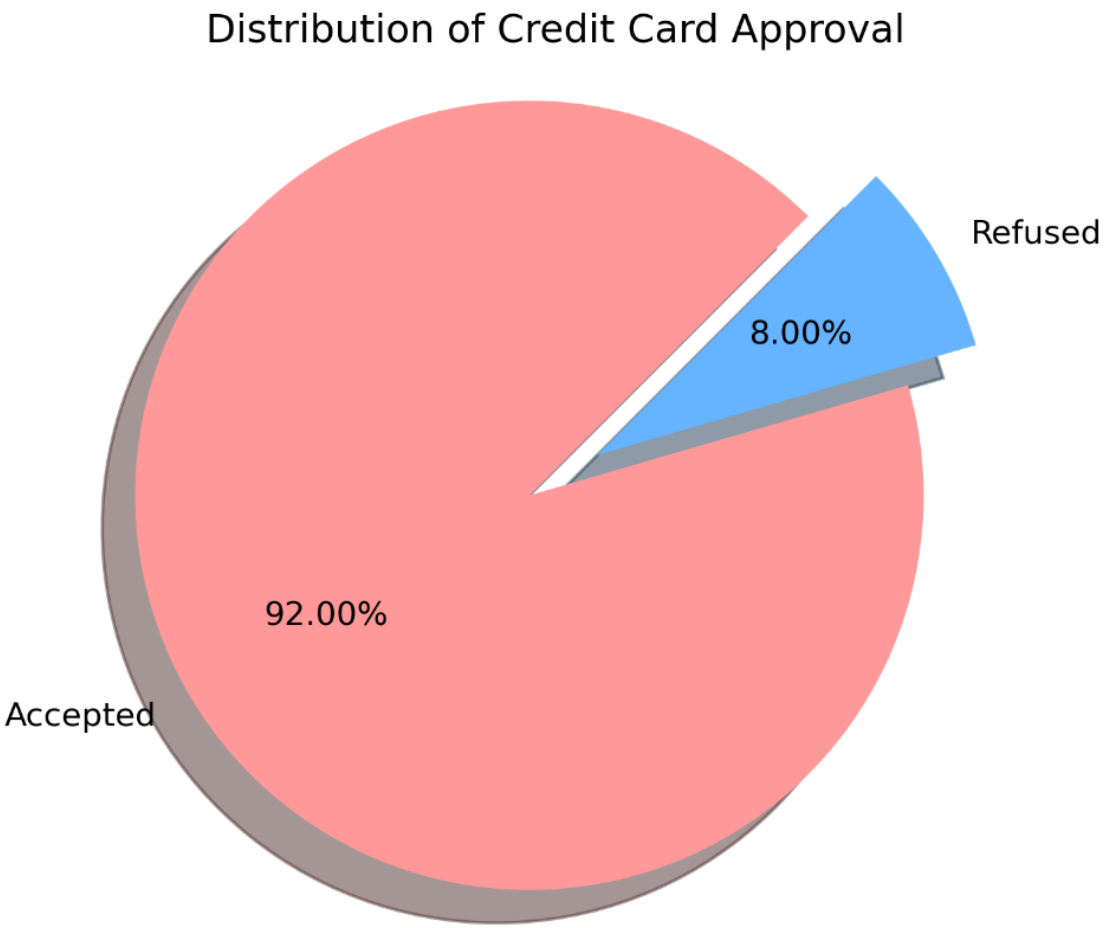


차트를 통해 확인할 수 있듯이 "0 Bad Debt"을 가진 고객들이 검토된 총 22,043개의 기록 중 87.7%를 차지합니다. 특히 이 그룹에 속한 모든 고객들이 신용 카드 승인을 받았으며, "0 Bad Debt 신용 카드 승인" 차트는 100% 승인 비율을 분명하게 보여줍니다.

데이터 불균형을 감소시키고 기계 학습 모델의 계산 시간을 줄이기 위해 "Total_Bad_Debt"가 0인 기록들을 제거하기로 결정했습니다.

데이터 전처리

1.변수 엔지니어링



탐색 데이터 분석' 부분에서, 우리는 '필터링'을 통해 데이터의 불균형 문제를 해결하기 위해 전문 지식을 적용하였습니다. 처음에는 '승인됨'과 '승인되지 않음' 레이블의 비율이 각각 99.51% 대 0.49%였습니다. 이러한 심각한 불균형을 인지한 후, 우리는 '승인됨' 레이블에서 특정 사례를 제거하여 균형 비율을 92%와 8%로 줄였습니다. 그러나 개선이 있었음에도 불구하고, 데이터의 불균형 문제가 완전히 해결되지는 않았습니다. 따라서, 우리는 SMOTE (Synthetic Minority Over-sampling Technique) 방법을 통한 재샘플링 기술을 적용하기로 결정했습니다. 이는 더 적은 데이터 그룹의 대표성을 강화하여, 후속 분석 및 모델링 과정을 위한 더 균형 잡힌 데이터 세트를 생성하는 것이 목표입니다.

실행한 '필터링' 과정은 데이터 불균형을 해결하기 위한 여러 특정 기준

을 포함하고 있습니다. 주로, Total_Bad_Debt가 0보다 큰 모든 고객을 제거하는 것을 포함했습니다. 또한, 학생이 아닌 고객들과 기타 특정 조건들도 배제했습니다. 이 엄격한 필터링 과정은 우리 데이터 세트의 균형을 효과적으로 변화시켰습니다. 결과적으로, '승인'과 '거절' 사례 간의 비율이 처음의 25,007:121에서 더 균형 잡힌 1,420:121로 크게 변했습니다.

영역 지식에 기반한 새로운 특징 생성:
Bad_Debt_to_Good_Debt_Ratio: 이는 총 나쁜 부채와 총 좋은 부채 사이의 비율입니다. 높은 값은 고객이 나쁜 신용 기록을 가질 수 있음을 나타낼 수 있습니다.
Age_to_Working_Years_Ratio: 이 비율은 고객의 나이와 그들이 일한 연수 사이의 관계를 보여줍니다. 낮은 비율은 고객이 어린 나이에 일하기 시작했을 수 있음을 나타내며, 높은 비율은 그들의 경력이 늦게 시작되었을 수 있음을 제안할 수 있습니다.

Income_Per_Family_Member: 이 변수는 가족 구성원당 평균 소득을 계산합니다. 이는 각 가정의 재정 능력에 대한 유용한 통찰을 제공할 수 있습니다.

다항 공학(Polynomial Engineering):
P.Total_Good_Debt = Total_Good_Debt**2
P.Total_Bad_Debt = Total_Bad_Debt**2

우리가 다항 특징 공학 기법을 사용하는 이유는 변수가 분류에 미치는 영향을 강화하기 위함입니다. 이는 모델이 더 복잡한 관계를 파악하는 데 도움을 줄 수 있으며, 따라서 F1 점수나 재현율을 크게 향상시킬 수 있습니다.

2. 가설 검정 및 특징 제거

T-검정 결과 'Owned_Mobile_Phone', 'Owned_Phone', 'Owned_Email', 'Owned_Work_Phone', 'Job_Title' 변수가 신용카드 승인에 유의미한 영향을 미치지 않는다는 강력한 통계적 증거가 없다고 밝혀졌습니다. (P-value > 0.05). 귀무가설을 기각할 수 없습니다.

처음에는 'Applicant_ID'와 같이 공헌하지 않는 것으로 간주된 변수들을 제외했습니다. 그 후의 탐색적 데이터 분석(EDA)과 가설 검정은 p-값과 t-통계를 이용하여 'Owned_Mobile_Phone', 'Owned_Phone', 'Owned_Email', 'Owned_Work_Phone', 'Job_Title'과 같은 변수들을 제외해야 한다고 확인했습니다. 이러한 예측 변수들의 제거는 모델 성능의 향상과 훈련 시간의 감소를 통해 그 제외를 정당화했으며, 예측 프레임워크에서의 제외를 뒷받침했습니다.

3. Box-cox 기술

The MEANS Procedure

Variable	Mean	Variance	Skewness	Kurtosis
Applicant_Age	38.7216093	81.3750457	0.2848098	-0.9984230
Years_of_Working	6.8118105	29.2723539	1.3754064	1.5924517
Total_Income	195733.06	10336936171	2.4378774	10.8720115
Total_Good_Debt	11.5600260	29.6829205	-0.0111484	-1.0093875
Total_Bad_Debt	2.7053861	17.8313281	6.0770772	49.7655032

분석 과정에서 Years_of_Working과 Total_Income 변수는 높은 왜도와 첨도를 나타내므로 Box-Cox 변환을 적용하여 비대칭성과 무거운 꼬리를 감소시킬 것을 고려해야 합니다. 반면에, Applicant_Age는 상대적으로 대칭적인 분포를 보이며 변환할 필요가 없습니다.

4. 데이터 정규화 및 특징 인코딩

이 전처리 단계에는 두 가지 주요 작업이 포함되었습니다: 숫자 데이터의 표준 스케일러를 사용한 정규화와 pd.get_dummies를 통한 범주형 변수의 인코딩입니다. 표준 스케일러는 숫자 특징을 표준화하여 평균이 0이고 단위 분산을 갖도록 함으로써, 많은 머신러닝 모델의 성능에 결정적인 역할을 합니다. 범주형 데이터의 경우, pd.get_dummies는 객체 타입 변수를 이진 열로 효과적으로 변환하여 모델이 이러한 범주를 효율

적으로 처리하고 해석할 수 있게 합니다.

BUILD PREDICT MODEL

예측 모델 개발 과정에서 우리는 AdaBoost, XGBoost 및 Logistic Regression과 같은 강력한 머신러닝 알고리즘을 사용할 것이며, 딥 러닝 기술은 사용하지 않을 것입니다. 이 결정은 우리 데이터 세트의 특성에서 비롯되었는데, 상대적으로 간단하고 크기가 제한적이어서 복잡성과 데이터 요구 사항이 있는 딥 러닝 모델보다는 전통적인 머신러닝 방법에 더 적합합니다.

1. 기준 모델

기본 모델 평가 단계에서, 우리는 여러 머신러닝 알고리즘을 통합한 파이프라인을 구축했습니다. 이 단계에서는 오직 정제된 데이터만 사용되었으며, 어떠한 고급 최적화 기술도 적용되지 않았습니다. 이 접근 방식은 각 모델이 데이터셋에서 기본적으로 얼마나 효과적인지 객관적으로 평가하기 위한 것으로, 후속 단계에서의 모델 선택 및 최적화를 위한 기초를 마련합니다.

```
AdaBoost Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.92      0.94       121
           1       1.00      1.00      1.00     25007

    accuracy                           1.00     25128
   macro avg       0.98      0.96      0.97     25128
weighted avg       1.00      1.00      1.00     25128

XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95       121
           1       1.00      1.00      1.00     25007

    accuracy                           1.00     25128
   macro avg       0.99      0.97      0.98     25128
weighted avg       1.00      1.00      1.00     25128

Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.58      0.72       121
           1       1.00      1.00      1.00     25007

    accuracy                           1.00     25128
   macro avg       0.97      0.79      0.86     25128
weighted avg       1.00      1.00      1.00     25128
```

모델 기준선을 구축하는 과정에서 우리는 원시 데이터 세트를 사용했습니다. 결과는 AdaBoost와 XGBoost 모델이 '거절'(즉, 신용 거부 사례)을 인식하는 데 있어 각각 94%와 95%의 높은 성능을 보여주어 다소 놀라웠습니다. 그러나 우리의 목표는 은행의 불필요한 손실을 방지하기 위해 '거절' 예측 능력이 더 높은 모델을 만드는 것입니다. 다음 단계에서는 머신러닝 모델에 데이터를 입력하기 전에 사용되는 기술들을 적용하여 모델의 예측 능력을 향상시킬 것입니다.

2. 모델에 기술 적용

교차 검증

모델 성능의 견고하고 편향되지 않은 평가를 보장하기 위해, 우리는 'n_splits=5'로 5개의 분할을 가진 StratifiedKFold 교차 검증 전략을 사용했습니다. 이 방법은 각 클래스의 샘플 비율을 보존하여, 각 훈련 및 테스트 폴드에서 원본 데이터셋의 클래스 분포를 유지합니다. 분할 전에 'shuffle=True'를 포함하는 것은 훈련 및 테스트 데이터셋의 무작위성과 다양성을 향상시켜, 특히 모델 과적합을 방지하고 모델의 일반화 가능성을 보장하는 데 유익합니다. 이 교차 검증 접근법은 우리 데이터셋의 클래스 불균형 맥락에서 신뢰할 수 있는 성능 지표를 달성하기 위해 중요합니다.

SMOTE

데이터셋의 클래스 불균형을 수정하기 위해, 우리는 합성 소수 클래스 오버 샘플링 기술(SMOTE)을 도입했습니다. 소수 클래스에 대한 합성 샘플을 생성함으로써, SMOTE는 더 균형 잡힌 클래스 분포를 달성하는 데 도움을 주며, 이를 통해 모델이 각 클래스의 중요성을 더 정확하게 반영하는 데이터셋에서 학습할 수 있도록 합니다. 이 기술은 데이터에서 종종 대표성이 낮은 소수 클래스에 대한 모델의 예측 성능을 향상시키는 데 특히 유용합니다.

주성분 분석

주성분 분석(PCA)은 데이터셋의 차원을 줄임으로써 우리 모델의 성능을 향상시키는 데 중요한 역할을 했습니다. 가장 변동성이 큰 주성분에 초점을 맞춤으로써, 우리는 필수 정보를 유지하면서 중복된 특성을 제거할 수 있었습니다. 이 차원 축소는 훈련 시간을 단축시키는 것뿐만 아니라 과적합 위험을 줄이고, 모델을 새로운 데이터에 대해 더 강인하고 일반화 가능하게 만드는 데에도 도움을 주었습니다.

기타 기술

앞서 언급한 클래스 가중치 전략 외에도, 데이터 전처리 섹션에서 자세히 설명한 바와 같이, 특히 '필터링'과 관련하여 모델 성능을 향상시키기 위한 일련의 기술들을 구현했습니다. 이에는 새롭고 유익한 속성을 생성하는 특성 공학, 가장 관련성 높은 예측 변수를 식별하는 특성 선택, 수치 데이터를 스케일링하는 정규화가 포함됩니다. 또한, 박스-콕스 변

환의 적용은 왜곡된 특성의 분산을 안정화하고 분포를 정규화하는 데 중요한 역할을 했습니다. 이러한 단계들을 종합적으로 적용함으로써, 우리 모델의 효율성과 예측력이 크게 향상되었습니다.

3. 개선

하이퍼파라미터 튜닝

우리는 RandomizedSearchCV를 사용하여 XGBoost 모델로 최적의 매개변수를 찾았습니다.

서브샘플: 각 트리가 데이터 샘플의 100%(1.0)로 훈련되도록 보장합니다.
정규화 람다(reg_lambda): 과적합을 피하기 위해 가중치에 L2 정규화 항을 1로 적용합니다.
정규화 알파(reg_alpha): 특징 선택을 위해 가중치에 0.1 값의 L1 정규화 항을 통합합니다.
추정기의 수: 앙상블에서 900개의 트리를 사용합니다.
최소 자식 가중치: 자식에 필요한 인스턴스 가중치의 합을 3으로 설정합니다.
최대 깊이: 각 트리가 최대 8 레벨까지 자랄 수 있도록 합니다.
학습률: 각 단계에서 보수적인 비율 0.01로 가중치를 조정합니다.
감마: 리프 노드에서 추가 분할을 위해 필요한 손실 감소를 최소화하여 감마 값 1을 사용합니다.
트리별 샘플링: 각 트리마다 90%의 특징을 사용합니다(0.9).

임계값 조정

우리는 모델이 거절된 고객을 식별하는 능력을 강화하기 위해 class 0의 재현율(recall)을 전략적으로 우선시하였습니다. 의사 결정 임계값을 0.49로 설정함으로써, 우리는 의도적으로 모델을 "거절" 클래스 쪽으로 편향시켰습니다. 이러한 미세한 조정은 모델이 거절 클래스를 효과적으로 식별하는 능력을 크게 향상시키는 동시에 정확도(accuracy)를 크게 희생하지 않도록 돕습니다.

Result

```
Fold 1 Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        25
           1       1.00      1.00      1.00       284

    accuracy                           1.00       309
   macro avg       0.98      1.00      0.99       309
weighted avg       1.00      1.00      1.00       309

Fold 2 Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        24
           1       1.00      1.00      1.00       284

    accuracy                           1.00       308
   macro avg       0.98      1.00      0.99       308
weighted avg       1.00      1.00      1.00       308

Fold 3 Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94        24
           1       1.00      0.99      0.99       284

    accuracy                           0.99       308
   macro avg       0.94      0.99      0.97       308
weighted avg       0.99      0.99      0.99       308

Fold 4 Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        24
           1       1.00      0.99      1.00       284

    accuracy                           0.99       308
   macro avg       0.96      1.00      0.98       308
weighted avg       0.99      0.99      0.99       308

Fold 5 Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        24
           1       1.00      1.00      1.00       284

    accuracy                           1.00       308
   macro avg       0.98      1.00      0.99       308
weighted avg       1.00      1.00      1.00       308
```

결과는 매우 고무적입니다. 우리는 모든 5개의 폴드에 걸쳐 "거절됨" 사례를 예측할 수 있는 모델을 성공적으로 구축했습니다. 이 일관된 성능은 클래스 0에 대한 거의 완벽한 재현율 점수에 반영되며, 신용 카드가 거절될 가능성이 있는 고객을 식별하는 모델의 견고함을 나타냅니다.

또한, 클래스 1에 대한 정밀도와 F1-점수도 높게 평가되며, 두 클래스 모두에 대해 높은 정확도를 유지하는 균형 잡힌 모델을 나타냅니다. 이러한 성능은 업계 표준을 충족할 뿐만 아니라 초과하여, 신용 카드 승인 절차의 일부를 자동화할 수 있는 신뢰할 수 있는 기반을 제공합니다.

IMPORTANT FEATURES

Rank	Pearson's Correlation	L1 Regularization	XGBoost
1	Total_Bad_Debt	P.Total_Bad_Debt	Bad_Debt_to_Good_Debt_Ratio
2	Bad_Debt_to_Good_Debt_Ratio	P.Total_Good_Debt	Total_Good_Debt
3	P.Total_Bad_Debt	Total_Good_Debt	Income_Type_3
4	Total_Good_Debt	Total_Bad_Debt	Income_Type_2
5	Total_Good_Debt	Bad_Debt_to_Good_Debt_Ratio	Income_Type_1
6	Age_to_Working_Years_Ratio	Housing_Type	Education_Type_1

이 분석에서는 중요한 변수, 특히 Pearson's Correlation, L1 Regularization 및 Ensemble.feature_importances_를 선택하기 위한 세 가지 방법을 사용합니다. 이어서, 이러한 방법론에 대한 평가가 수행될 것입니다.

피어슨의 상관관계
특징 선택에 활용되는 이 방법은 목표 변수와 가장 높은 절대 상관 관

계를 나타내는 특징을 식별하는 데 효과적입니다. 이는 필터 방법으로 작동합니다. 즉, 다른 기능과의 상호 작용을 고려하지 않고 각 기능을 독립적으로 평가합니다.

L1 정규화

이러한 형태의 회귀는 높은 수준의 다중 공선성을 나타내는 모델이나 변수 선택이나 매개변수 제거와 같은 모델 선택의 특정 측면을 자동화해야 하는 시나리오에 특히 유리합니다.

Pearson's Correlation, Lasso, XGBoost의 세 가지 특징 중요도 평가 방법을 통해 얻은 결과를 기반으로 합니다.
모델 - 이 보고서는 모델에서 신용 카드 승인을 분류할 때 각 변수의 기여도를 분석합니다.

분석에서는 "Total_Bad_Debt", "Total_Good_Debt" 및 "Income_Type_3"을 포함한 여러 변수가 상당히 중요하다고 강조합니다. 특히, Feature Engineering 과정을 통해 "Bad_Debt_to_Good_Debt_Ratio", "P.Total_Bad_Debt"와 같은 추가 변수가 중요한 것으로 나타났습니다. 및 "P.Total_Good_Debt"입니다.

제한 사항 이해

이 연구는 신용카드 승인 위험 평가를 개선하는 데 중요한 진전을 이루었지만, 몇 가지 중요한 제약 사항을 인식하는 것이 필수적입니다.

이 연구는 기본 고객 정보만을 포함하는 데이터셋에 의존하고 있습니다. 모델의 예측 능력을 강화하기 위해서는 더 구체적이고 포괄적인 데이터를 수집하는 것이 필수적이며, 이를 통해 복잡하고 강력한 딥 러닝 모델의 활용이 가능해집니다.

감사의 말

시간을 내어 우리 팀의 보고서를 읽어주셔서 진심으로 감사드립니다. 본 연구가 진행되는 동안 지지와 도움을 주신 모든 분들께 진심으로 감사드립니다. 이번 대회에 참가할 수 있도록 지도와 격려를 해주신 교수님들께 특별한 감사를 드립니다. 귀하의 통찰력과 전문 지식은 우리에게 매우 귀중한 것이었습니다.

참조 소스

https://dl.ucsc.cmb.ac.lk/jspui/bitstream/123456789/4593/1/2018%20BA%20026.pdf
https://publications.vtt.fi/julkaisut/muut/2006/customer_churn_case_study.pdf
https://ieeexplore.ieee.org/abstract/document/8935884
https://medium.com/@amansangal9/predicting-credit-card-approvals-8409c5280f91
https://www.datacamp.com/projects/558