AWS 테블러 데이터 준비 (탐색, 피쳐 선택 등)

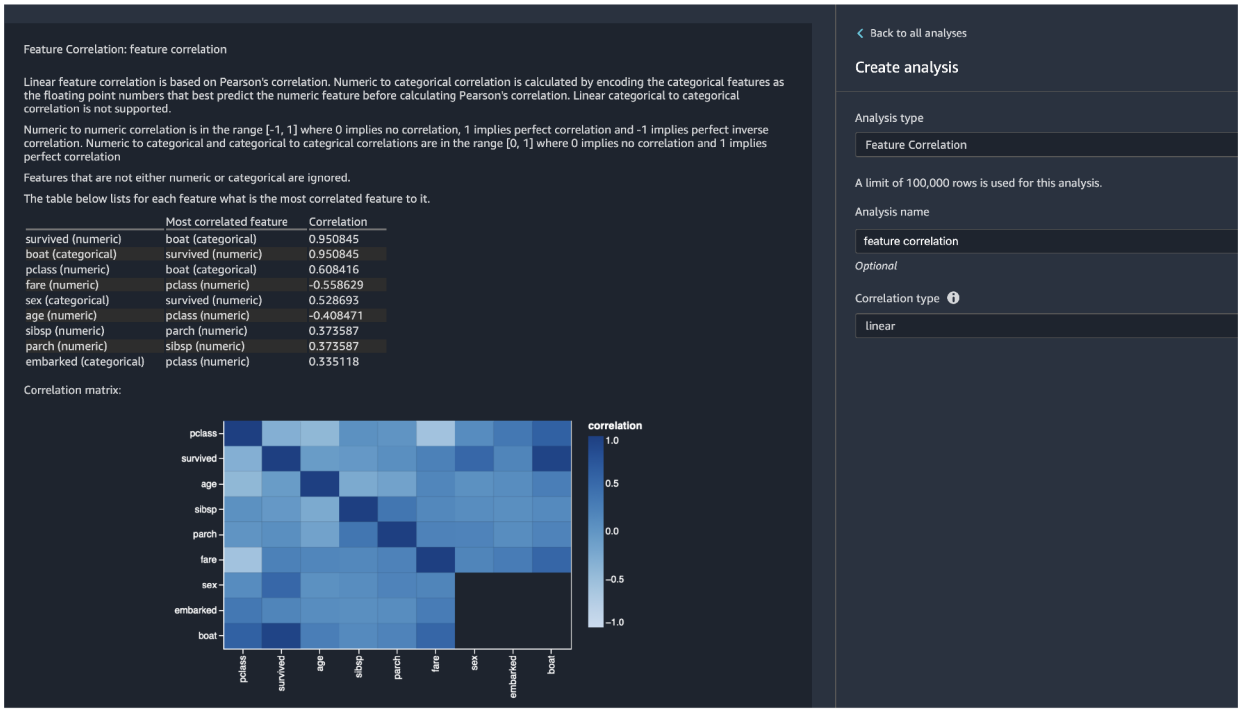
기본적으로 SageMaker Data Wrangler 와 AutoPilot의 데이터 탐색 및 데이터 피쳐 변환의 자동화 툴을 사용해보는 것을 권장 드립니다.
이렇게 데이터를 돌리고 결과를 보고 나서, 필요하면 수동으로 코드를 짜서 해보는 접근 방법을 권장 드립니다. (아래 3. 수동으로 코딩하여 데이터 탐색 참조 바람)

최근에 Data Wrangler, AutoPilot 의 기능이 많이 향상이 되었습니다.
필요한 기능을 아래에서 확인 하시면 좋겠습니다.

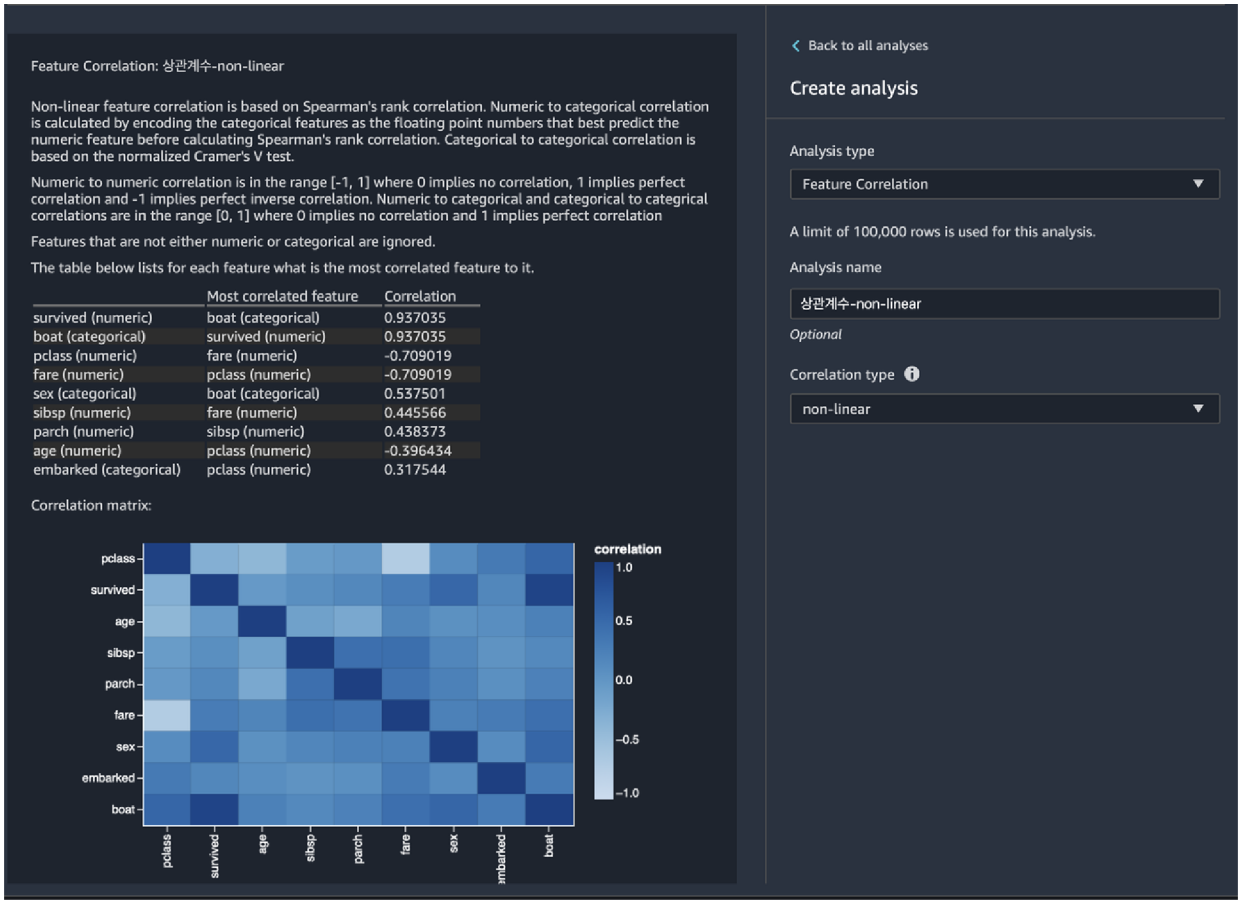
1. SageMaker Data Wrangler

주요 특징 스크린 샷

(linear - 피어슨 상관계수 이용) : 상관계수 분석을 통하여 유사도(피어슨 계수)가 높은 것의 피쳐를 제거 할 수 있음.

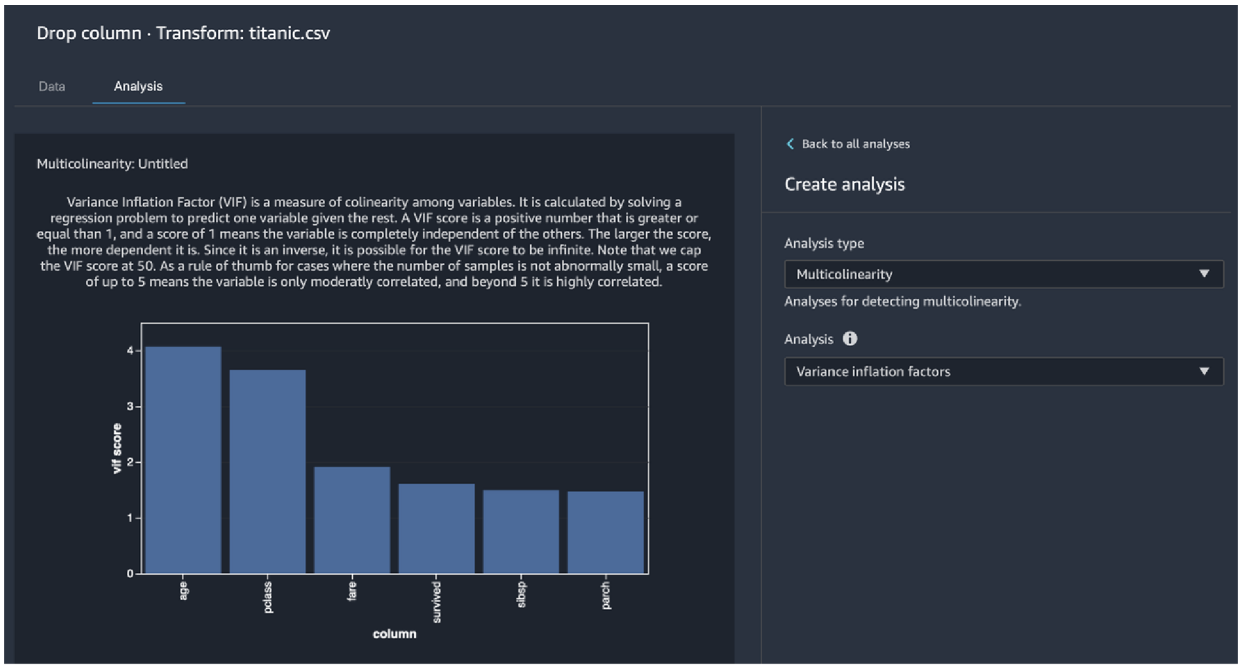


(Non-Linear - Spearman's Rank 상관계수 이용) : 상관계수 분석을 통하여 유사도 높은 것의 피쳐를 제거 할 수 있음.

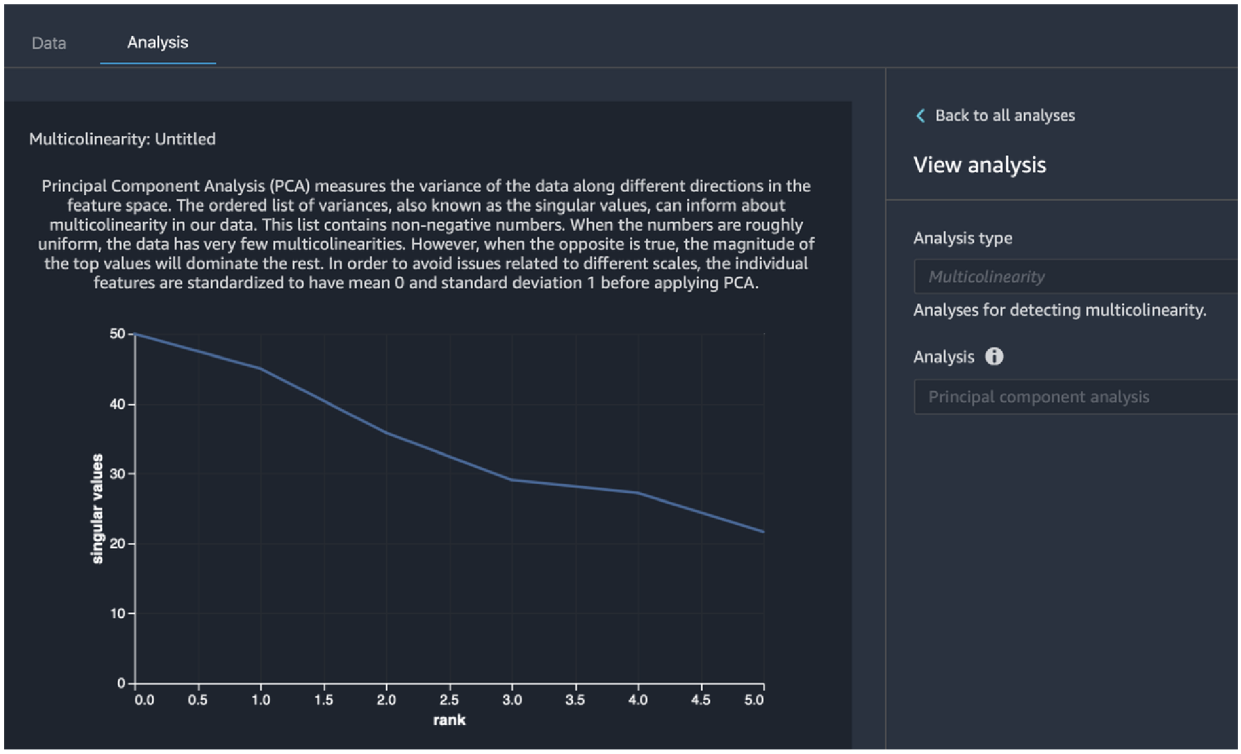


VIF (Variance Inflation Score) 를 통한 비슷한 피쳐를 검출

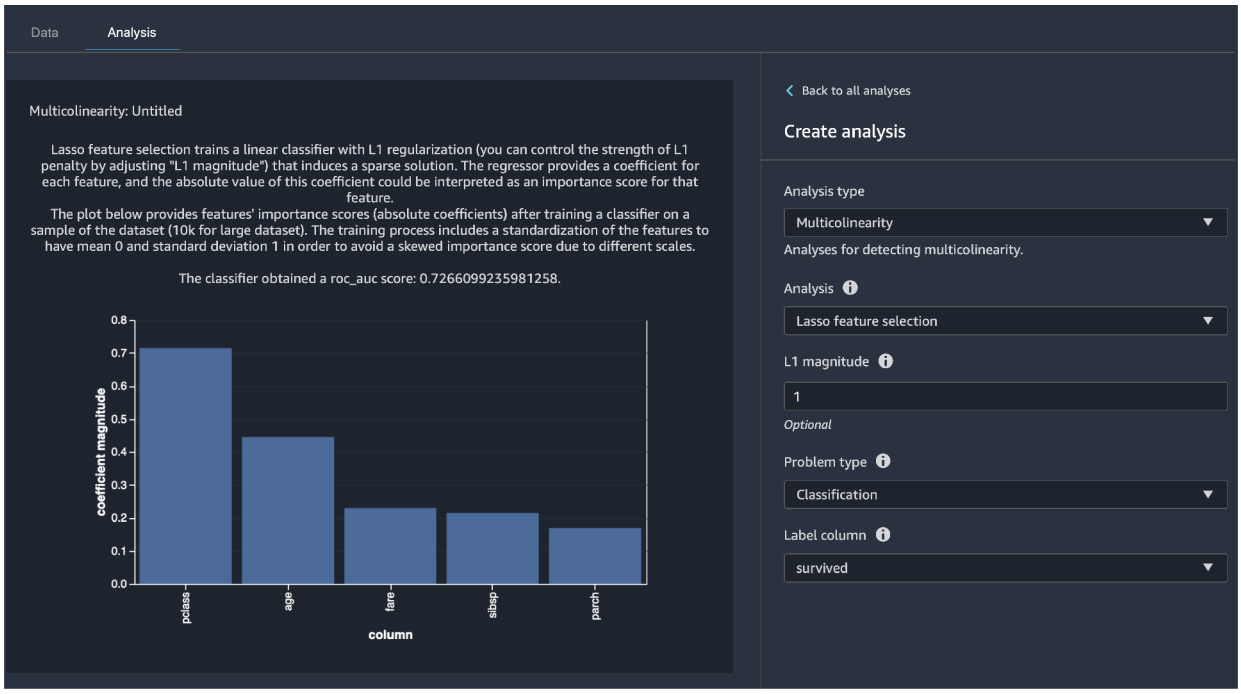
- 참고: What is a Variance Inflation Factor?
  - https://www.statisticshowto.com/variance-inflation-factor/



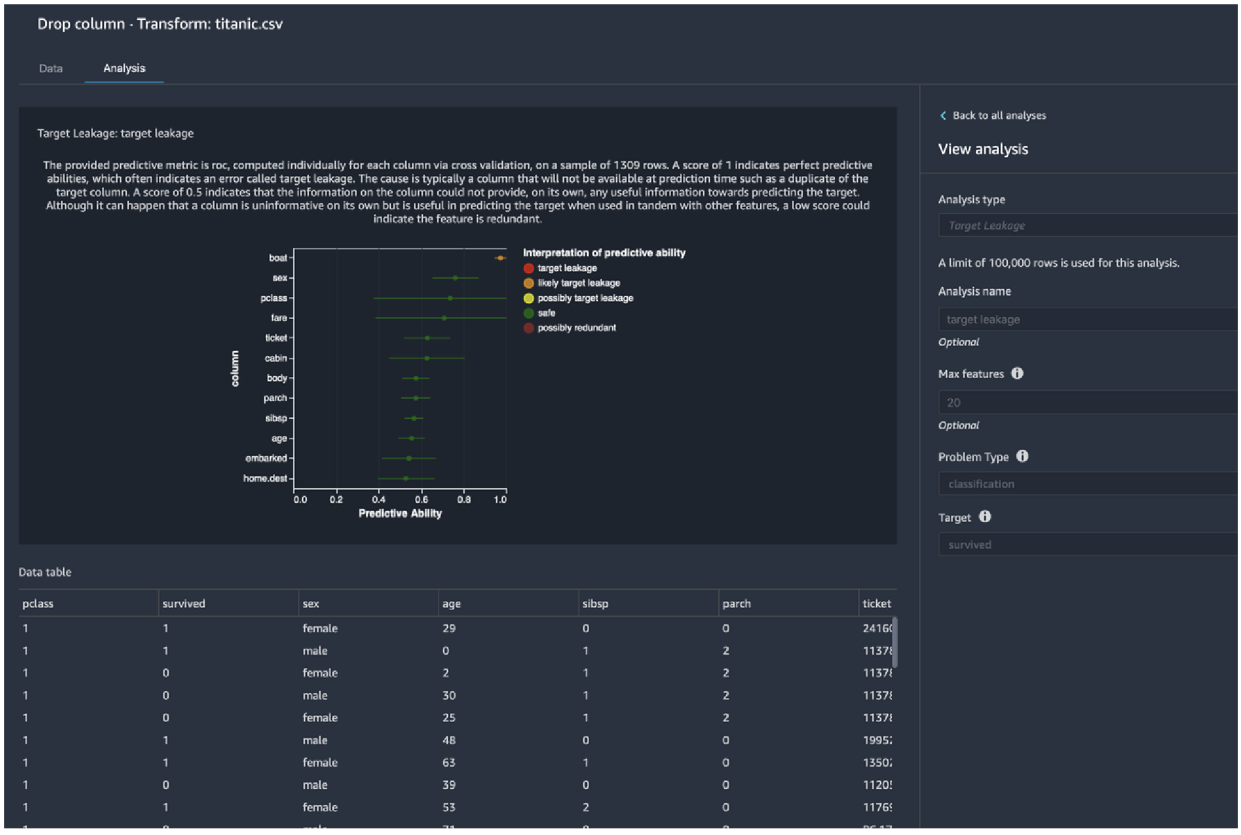
PCA 분석을 통한 전체적인 비슷한 치쳐 정도 파악



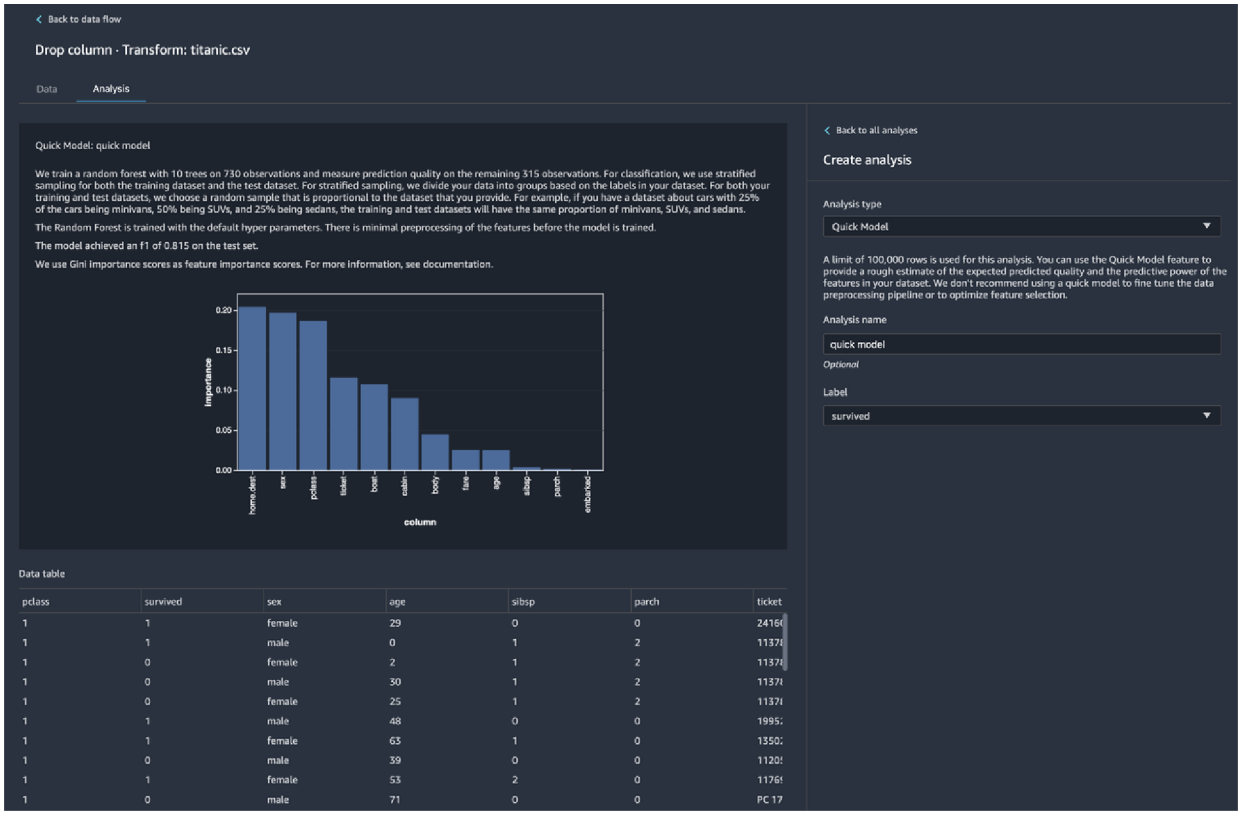
라소 피쳐 선택: Lasso Classifier 를 실행하여 중요한 피쳐를 구별하고, 그렇지 않은 것을 제거할 수 있음.



타겟 리키지 현상 확인 및 피쳐 중요성: 레이블의 정답이 누출 되는 것을 감지.



Qucik 모델: 간단히 빠르게 모델을 생성하여 평가 지표 및 피쳐 중요성을 봄.



공식 서비스 페이지 및 개발자 가이드

Amazon SageMaker Data Wrangler

- https://aws.amazon.com/sagemaker/data-wrangler/

Prepare ML Data with Amazon SageMaker Data Wrangler

- https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html

블로그:

[Dec 2020] Introducing Amazon SageMaker Data Wrangler, a Visual Interface to Prepare Data for Machine Learning

- https://aws.amazon.com/blogs/aws/introducing-amazon-sagemaker-data-wrangler-a-visual-interface-to-prepare-data-for-machine-learning/

[Dec 2020] Exploratory data analysis, feature engineering, and operationalizing your data flow into your ML pipeline with Amazon SageMaker Data Wrangler

- https://aws.amazon.com/blogs/machine-learning/exploratory-data-analysis-feature-engineering-and-operationalizing-your-data-flow-into-your-ml-pipeline-with-amazon-sagemaker-data-wrangler/

[Sep 2021] Schedule an Amazon SageMaker Data Wrangler flow to process new data periodically using AWS Lambda functions

- https://aws.amazon.com/blogs/machine-learning/schedule-an-amazon-sagemaker-data-wrangler-flow-to-process-new-data-periodically-using-aws-lambda-functions/

[강력 추천] [Nov 2021] Accelerate data preparation using Amazon SageMaker Data Wrangler for diabetic patient readmission prediction

- https://aws.amazon.com/blogs/machine-learning/accelerate-data-preparation-using-amazon-sagemaker-data-wrangler-for-diabetic-patient-readmission-prediction/

2. SageMaker AutoPilot

주요 특징 스크릿 샷

오토 파일럿을 실행하고 데이타 요약을 확인

Dataset Summary

Dataset Properties

Rows	Columns	Duplicate rows	Target column	Missing target values	Invalid target values	Detected problem type
4000	19	0.00%	Churn?	0.00%	0.00%	BinaryClassification

Detected Column Types

	Numeric	Categorical	Text	Datetime	Sequence
Column Count	18	0	0	0	0
Percentage	100.00%	0.00%	0.00%	0.00%	0.00%

레이블인 타겟의 이상치 및 권고 사항 예시

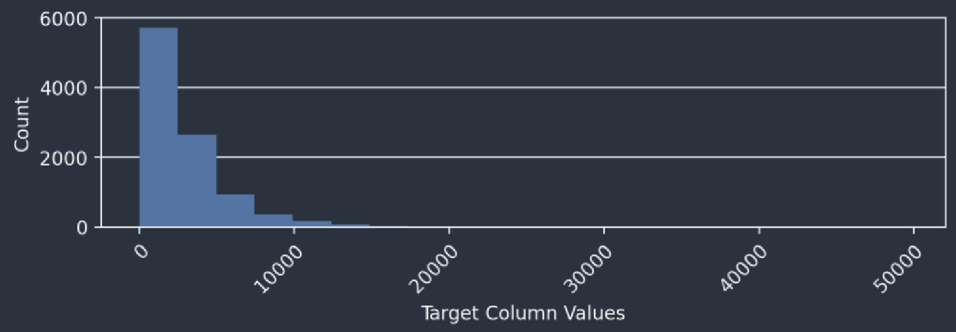
Target Analysis

⚠ High severity insight: "Heavy tailed target"

The distribution of values in the target column is heavy tailed and might contain outliers. As the outliers induce high errors when optimizing MSE (or similar loss functions) ML algorithms tend to focus on them when training the model. That might result in sub-par prediction quality for the non-outlier rows. If it is important to predict the extreme target values well, then there might be no need for further action. If prediction of extreme values is not important, consider clipping extreme target values. Clipping or removing outliers can be done with Amazon SageMaker Data Wrangler using the "Robust standard deviation numeric outliers" transform under "Handle outliers".

The column **y** is used as the target column. See the distribution of values (labels) in the target column below:

Mean	Median	Minimum	Maximum	Skew	Kurtosis	Number of Uniques	Outliers Percentage	Invalid Percentage	Missing Percentage	Missing Count
3017.90	2116.24	0.67	121012.25	2.86	16.33	130809	1.30%	0.00%	0.00%	0



Histogram of the target column values. The orange bars contain outliers and the value below them is the outliers average.

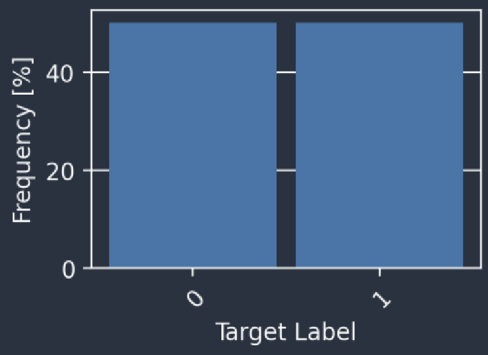
레이블 값인 타겟을 확인

Target Analysis

The column **Churn?** is used as the target column. See the distribution of values (labels) in the target column below:

Number of Classes	Invalid Percentage	Missing Percentage
2	0.00%	0.00%

Target Label	Frequency Percentage	Label Count
0	50.05%	2002
1	49.95%	1998



Histogram of the target column labels.

상관계수 분석

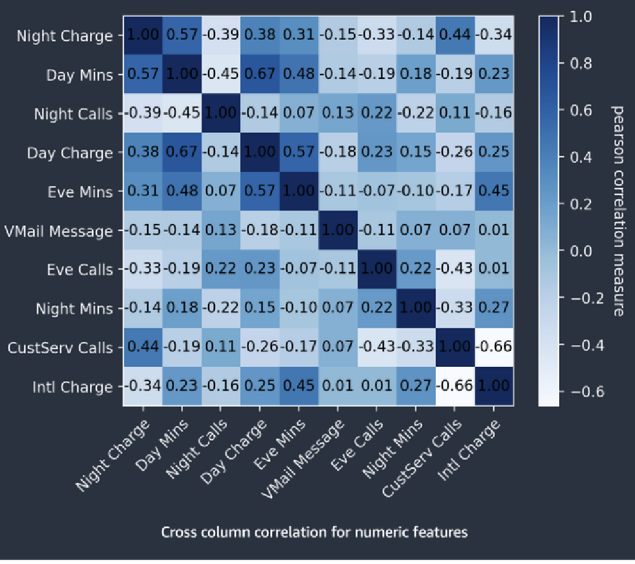
Duplicate Rows

No duplicate rows were found when testing a random sample of 4000 rows from the dataset.

Cross Column Statistics

Amazon SageMaker Autopilot calculates Pearson's correlation between columns in your dataset. Removing highly correlated columns can reduce overfitting and training time. Pearson's correlation is in the range [-1, 1] where 0 implies no correlation, 1 implies perfect correlation, and -1 implies perfect inverse correlation.

The full correlation matrix between the 10 most predictive numeric features is presented below.



Cross column correlation for numeric features

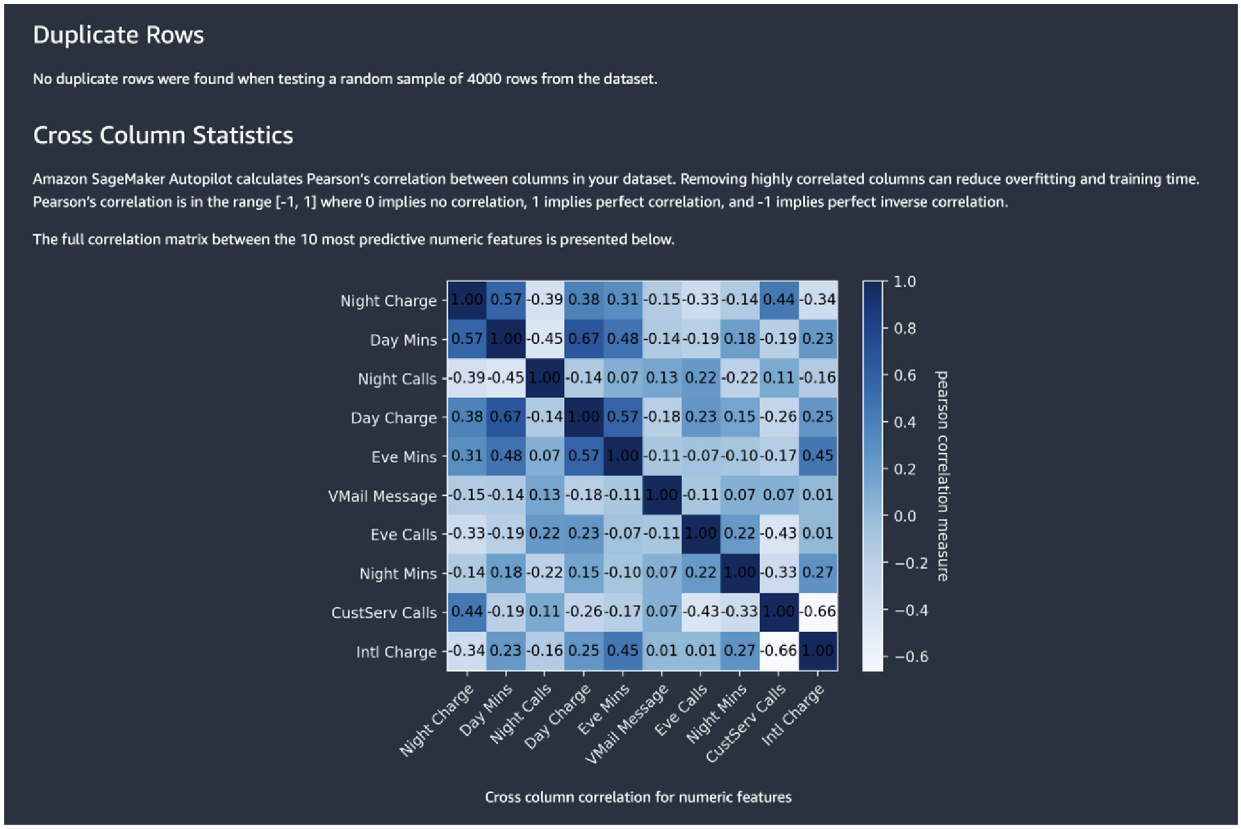
주어진 데이터에서 아웃라이어 데이터 샘플 출력

Anomalous Rows

Anomalous rows are detected using the Isolation forest algorithm on a sample of **4000** randomly chosen rows after basic preprocessing. The isolation forest algorithm anomaly score to each row of the dataset it is trained on. Rows with negative anomaly scores are usually considered anomalous and rows with positive anomaly score considered non-anomalous. When investigating an anomalous row, look for any unusual values - in particular any that might have resulted from errors in the gatherin processing of data. Deciphering whether a row is indeed anomalous, contains errors, or is in fact valid requires domain knowledge and application of business logic.

Inspect the rows below, to see if any of those are anomalous. A subset of rows is presented below. Anomaly score is presented as the left most column; Smaller values higher chance that the row is anomalous.

	Anomaly Scores	Account Length	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Night Mins	Night Calls	Night Charge	Intl Min
1105	-0.135268	11	300	4.526399619684806	6	5.4528933196961935	8.522019642641109	12	3.5636554576027586	300	0.6046913918959254	6.89620
1044	-0.107473	171	400	1.9576360316901276	3	9.913322488589957	13.622096683112519	11	3.4000000594176942	350	1.3570973191027529	5.29350
1889	-0.101399	189	600	0.873321480510346	1	5.3019420650603095	1.9765462179089448	6	7.945452426267903	150	0.10718975162271073	4.04664
1293	-0.092951	24	500	10.449854912097138	5	7.818642848211213	5.4261277920375095	7	9.139956354980876	50	2.345055816886982	4.01314
1100	-0.092276	200	400	15.284420479859804	2	9.908326501631505	9.261201354953398	1	4.118486851196563	150	4.639671283314684	5.12638
3476	-0.089575	96	0	7.0597196958841275	3	2.1812053134807368	5.269741669979053	0	8.19337962132131	100	3.5062591172266298	2.89700
2387	-0.089259	150	0	2.438635146313733	2	5.815546964993327	2.664652098973234	11	9.253883768794806	250	0.908555272927808	4.62631
2604	-0.088420	200	0	2.688124983469433	2	10.028018224399833	10.902095375944116	13	5.392957206120896	300	2.7596400415550395	4.10194
3718	-0.086283	4	100	14.000023345763696	3	10.535626854235836	7.868834044870532	8	6.891634226620591	200	4.626493523379646	3.42812
505	-0.085814	41	0	16.07060659616391	3	12.17417236521136	8.09160772616227	9	5.406542352487855	150	6.135542276680621	6.44604



공식 서비스 페이지 및 개발자 가이드

Amazon SageMaker Autopilot

- https://aws.amazon.com/sagemaker/autopilot/

Automate model development with Amazon SageMaker Autopilot

- https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html

블로그

[Nov 2021] Use integrated explainability tools and improve model quality using Amazon SageMaker Autopilot

- https://aws.amazon.com/blogs/machine-learning/use-integrated-explainability-tools-and-improve-model-quality-using-amazon-sagemaker-autopilot/

3. 수동으로 코딩하여 데이터 탐색 자료:

Tablur Data (CSV 데이터) 피쳐 선택 기본 가이드

- https://github.com/gonsoomoon-ml/Self-Study-On-SageMaker/blob/main/data_preparation/Feature_Selection_Guide.md

ML 데이터 준비 및 ML Workflow 프로토 타이핑

- https://github.com/gonsoomoon-ml/ml-data-prep-workshop