기계학습 알고리즘 기반 항공편 지연 분류 예측 분석

[1]홍준석, [2]이재은, [3]권민지, [1]김동원, *조석헌
[1]건국대학교, [2]연세대학교, [3]조선대학교, *University of California, San Diego (UCSD)

acacia0522@gmail.com, lje1203@yonsei.ac.kr, knj23811906@gmail.com, master@dwer.kr, *justinshcho@gmail.com

Analysis of Aircraft Delay Prediction based on Machine Learning Algorithms

[1]Junseok Hong, [2]Jaeeun Lee, [3]Minji Kwon, [1]Dongwon Kim, *Seokheon Cho
[1]Konkuk University, [2]Yonsei University, [3]Chosun University, *University of California, San Diego (UCSD)

요 약

This study aims to propose a flight delay prediction model to minimize personal and national economic losses due to unexpected flight delays caused by continuous time-varying weather conditions and to ensure efficient flight schedule operations. We utilized departure flight data from Chicago O'Hare International Airport, known for its high number of flights and departure delay rates, as well as weather data collected at the airport. Due to the imbalance in the dataset used for the flight delay classification prediction model, the downsampling and Synthetic Minority Over-sampling Technique (SMOTE) methods were employed to balance between the majority class of non-delayed flights and the minority class of delayed flights. Additionally, Logistic Regression and Random Forest algorithms were considered for the prediction model. Analysis for our provided aircraft delay prediction models showed that the RF-based model, with SMOTE applied to the dataset including weather data, exhibited the best performance. Furthermore, we introduced the necessity of using the average of the results from training the data separated by different seasons rather than training the entire year's data without separation by season to improve performance of prediction for the minority class of actual delayed flights.

Ⅰ. 서론

미국 연방항공청 Federal Aviation Administration (FAA) 행정부는 2019년 기준 항공편 지연으로 인해 연간 약 330억 달러의 비용이 발생할 것으로 예측했다 [1]. 예기치 못한 항공편의 지연은 경제적으로 큰 손실을 야기할 뿐만 아니라, 탑승객들이 불편함을 느끼고 항공사에 대한 신뢰도 또한 잃을 수 있다. 따라서 본 연구에서는 항공편 지연 관련 데이터를 기반으로 항공사와 탑승객들이 항공편을 원활하게 이용할 수 있도록 항공편 지연 예측 모델을 제공하고자 한다. Bureau of Transportation Statistics (BTS)에서 제공하는 여러 공항들의 데이터에 따르면, 2018년을 기준으로 Chicago O'Hare 국제공항 (ORD)에서 출발하는 항공편들의 총 지연 시간과 취소 항공편이 각각 68,000시간과 7,000건으로 미국의 모든 공항에서 가장 높게 나타났다 [2]. 따라서 본 연구에서는 BTS 데이터를 이용해 ORD 공항에서 항공편 지연 분류 예측 모델을 제시하고자 한다.

Sun Choi *et al.* 는 BTS 데이터 세트를 사용하여 항공편 지연 정보에 대한 분석을 수행하였다 [3]. 분석 결과 총 항공편의 20%가량이 지연되고 있었고, 지연의 40%정도가 날씨의 영향을 받는 것으로 결론지었다. 전처리 과정으로 데이터의 불균형 문제를 Synthetic Minority Over-sampling Technique (SMOTE) 기법을 사용하여 해결하였다. 그 결과, Random Forest (RF)의 정확도 (accuracy)가 81.37%로 가장 높았다. 하지만, 지연관련 항공편 데이터는 불균형 데이터 세트이므로, 정확도 이외에 F1 score 또는 재현율 (recall)과 같은 불균형 데이터를 대상으로 하는 모델들의 성능 지표들 또한 확인해야 한다. 왜냐하면, 특히나 관심을 가지는 지연된 항공편에 대하여 올바르게 예측하는 모델 제시가 더욱 중요하기 때문이다. Seongeun Kim *et al.* 는 Incheon airport (ICN), John F. Kennedy airport (JFK) 및 Midway airport (MDW) 등의 세 공항에서 수집한 10년치 데이터를 사용하여 항공편의 지연 예측을 수행하였다 [4]. 해당 연구 결과 전반적으로 RF에서 가장 우수한 성능을 보였다. JFK 공항을 기준으로 0.852의 정확도, 0.882의 재현율 그리고 0.856의 F1 score들의 모델 성능 결과를 얻었다. 이러한 모델 성능 결과 자체는 우수하다고 평가할 수 있지만, 항공기들의 출발 이후의 정보인 항공기 실제 출발 시간 (Actual departure time)이 출발 지연 예측에 사용되고 있음을 확인할 수 있었다. 해당 변수를 이용하는 것은 미래 시점의 데이터를 사용한 점과 종속변수의 값을 직접적으로 얻을 수 있다는 점에서 부적절하다고 판단하였다.

따라서 본 연구에서는 항공기 출발 지연 시간을 직접적으로 유추할 수 있는 특성들을 항공편 지연 분류 예측 시 배제시켜 위와 같은 오류를 방지하였다. 4계절이 분명하게 존재하는 ORD 공항에서 계절에 따른 항공편 지연 관계를 분석하기 위해서 분기별 날씨 특성에 대한 항공편 지연 예측을 수행하고자 한다. 추가적으로 항공편 지연에 날씨 관련 특성들이 미치는 영향을 확인하고자 날씨 변수의 유무에 따른 예측 모델 결과를 비교하였다.

본 논문의 구성은 다음과 같다. 제 2장은 본 연구에서 사용하는 데이터 세트에 대한 전처리 과정, 불균형 데이터 해소를 위한 다양한 샘플링 기법 그리고 모델에 적용한 기계학습 알고리즘들에 대해 설명한다. 제 3장은 전처리 과정을 거쳐 구축한 데이터 세트들에 기계학습 알고리즘들을 적용하여 얻은 예측 성능에 대한 결과를 분석한다. 최종적으로, 제 4장에서 본 연구의 결과와 향후 연구 과제를 제시하며 마무리 짓고자 한다.

Ⅱ. 데이터 전처리 과정 및 데이터 불균형 해소

2.1 데이터 세트 전처리 과정

본 연구에서는 Bureau of Transportation Statistics (BTS)를 통해 2015년 01월 01일부터 2019년 12월 31일까지 총 5년간의 ORD 국제공항에서 출발하는 항공편 관련 데이터를 이용하였다. 해당 데이터는 총 1,496,636편의 항공편 정보를 담고 있었고, 그 중 불필요한 데이터와 결측치를 삭제하는 전처리 과정을 통해 모델의 예측 정확도를 높이고자 하였다. 우선 109개의 변수들 중 중복된 정보를 갖는 변수들을 제거하고 하나로 병합하는 작업을 수행하였다. 추가적으로 지연 운행과 관련된 변수들 중 취소 (cancelled) 또는 우회 (diverted) 변수를 통해 항공편들의 취소 또는 우회 운항한 사실을 확인할 수 있다. 이 두 변수들의 데이터 샘플은 총 데이터의 2% 미만에 불과했고, 취소 또는 우회 운항한 항공편들에 해당하는 나머지 변수들은 결측치로 기록되어 있기 때문에 모델 학습에 방해가 될 요소로 판단하여 두 변수들을 삭제하였다.

날씨 데이터 세트는 Visual Crossing에서 제공하는 데이터 세트를 가공하여 사용한다 [5]. 항공편 데이터 세트와 동일하게 2015년부터 2019년까지 ORD 국제공항 날씨 데이터를 한 시간 단위로 수집하였다. 날씨 데이터 세트에서는 체감온도 (feelslike), 강수확률 (precipprob), 태양에너지 (solarenergy), 일출 시간 (sunrise), 일몰 시간 (sunset) 그리고 기상 데이터 측정지 식별 번호 및 위치 (stations)와 같이 항공편 지연에 영향이 없다고 판단한 변수들을 삭제하였다.

최종적으로 두 데이터 세트를 하나로 병합하기 위해, 항공편 데이터 세트를 한 시간 단위로 기록된 날씨 데이터 세트에 맞게 가공하는 작업을 수행하였다. 분석의 편의를 위해서 항공편 데이터 세트에서 항공기마다 예정 출발 시간을 기반으로, 출발 시간의 분 단위가 30분 이상인 경우 올림하고 30분 미만인 경우 내림 연산을 적용하여 모든 항공기들에 대해서 예정 출발 시간을 한 시간 단위로 수정하였다. 예를 들면, 아침 8시 32분이 출발 예정이었던 항공기의 수정된 출발 예정 시간은 9시 정각이 되는 것이다. 이후, 두 개의 항공편 그리고 기상 데이터 세트를 시간에 맞춰 병합하는 작업을 수행하여 하나의 데이터 세트를 구축하였다.

본 연구는 날씨의 영향을 고려하는 것뿐만 아니라, 분기별 날씨 특성에 따른 항공편 지연 여부도 분석한다. 이후 데이터 세트를 1Q (1~3월), 2Q (4~6월), 3Q (7~9월) 그리고 4Q (10~12월)로 총 4분기로 분리하여 계절 특성을 고려하였다. 이로써, 1Q, 2Q, 3Q, 4Q 그리고 전체 연도에 해당하는 5개의 데이터 세트가 구축되는 것이다. 또한, 날씨가 항공편 지연에 미치는 영향을 확인하기 위해 날씨 변수를 제거한 데이터 세트를 항공편 지연 분류 예측 시 추가적으로 고려하였다. 즉, 날씨 관련 변수들을 고려한 5개의 데이터 세트들 그리고 고려하지 않은 5개의 데이터 세트들을 종합하여 총 10개의 데이터 세트를 생성하여 예측 모델을 분석하였다.

그림 1 은 항공편 지연 분류 예측 모델의 종속 변수인 항공편 지연 여부에 대한 데이터 비율을 나타내고 있다. 항공편에 대한 지연 기준은 BTS에서 제시하는 15분과 동일하게 적용하였다. 즉, 출발 예정 시간에 비교하여 실제 출발 시간이 14분 이하이면 해당 항공편은 지연되지 않은 (non-delayed) 라벨로 분류되고 15분 이상이면 해당 항공편은 지연된 (delayed) 라벨로 분류한다. 그림 1에서 확인할 수 있는 것처럼 지연된 항공편과 지연되지 않은 항공편은 비율은 각각 21.6%와 78.4%로 불균형 데이터 세트이다. 본 연구에서 제시하는 항공편 지연 분류 예측 모델은 이진 분류 (binary classification) 모델에 해당한다.

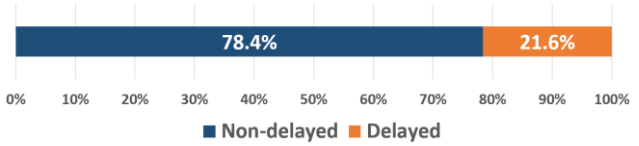


그림 1. 종속 변수 데이터 비율

2.2 불균형 데이터 해소를 위한 샘플링 기법

불균형 데이터 세트는 소수 클래스에 대한 학습이 정상적으로 이루어지지 않아 분류에 큰 오차가 발생할 수 있다 [6]. 따라서 불균형 문제를 해결하기 위해 본 연구에서는 down sampling과 oversampling 기법중의 하나인 Synthetic Minority Over-sampling Technique (SMOTE)와 같은 두 샘플링 기법을 적용하여 예측 모델들을 분석하였다. Down sampling은 다수 클래스의 샘플 수를 무작위로 제거함으로써 소수 클래스와의 분포를 균형 있게 만든다. 본 연구에서는 소수 클래스인 항공편 지연에 대한 예측을 시도하기 때문에, downsampling을 통해 모델이 소수 클래스에 더 학습할 수 있게 만들어 예측 모델의 성능을 높이고자 하였다. SMOTE는 소수 클래스의 샘플을 합성하여 늘리는 기법이다. 이는 새로운 샘플을 생성하기 때문에, 모델이 특정 샘플에 과적합 되는 것을 방지하고 더 다양한 패턴을 학습할 수 있도록 하는 것이다.

2.3 학습 알고리즘 및 성능 지표

본 연구에서는 Logistic Regression (LR)과 Random Forest (RF)의 두 알고리즘들을 사용하여 항공편 지연 분류 예측 모델의 성능을 평가 및 비교하였다.

LR은 입력 특성의 선형 조합을 사용하여 종속 변수의 발생 확률을 추정한다 [7]. 종속 변수는 이항 분포를 따르며, 독립 변수와 종속 변수 간의 관계를 설명하는데 사용된다. 각각의 입력 변수에 가중치를 곱해 합한 값을 사용한다. RF는 여러 개의 의사 결정 트리를 생성하고, 각 트리가 데이터 세트의 서로 다른 무작위 샘플로부터 독립적으로 학습한다 [8]. 각 트리는 다른 트리와 조금씩 다른 관점을 데이터에 적용하고, 모델 전체의 다양성을 증가시키며 학습을 진행한다. 따라서 일반화된 학습을 진행하여 과적합을 방지한다.

예측 모델들의 성능을 분석하기 위해서 고려하는 성능 지표들로는 정확도 (accuracy), 재현율 (recall), F1 score 및 ROC-AUC들이다. 4개의 성능 지표들 모두 0과 1 사이의 값을 가질 수 있으며, 1에 가까울수록 해당 모델의 성능이 높아진다고 할 수 있다.

Ⅲ. 항공편 지연 분류 예측 모델 분석

날씨 변수의 유무와 계절 특성을 고려한 10개의 데이터 세트에 downsampling과 SMOTE 기법을 적용하였다. 2015년부터 2018년까지의 데이터 세트를 학습 데이터 세트로 2019년 데이터 세트를 테스트 데이터 세트로 나누어서 모델 학습 및 테스트를 진행하였다.

그림 2는 항공편 지연 분류 예측 모델들의 다양한 성능을 보여주고 있다. 모든 성능 지표들을 전반적으로 살펴보았을 때, 기상 관련 데이터가 포함된 데이터 세트에 SMOTE를 적용한 RF 기반 모델의 성능이 다른 모델들의 성능보다 향상됨을 확

인할 수 있다. 가장 성능이 우수한 상기 경우의 모델에 있어서, 2019년을 분기별로 분리한 데이터 세트들에 대한 예측 모델들 (Case I)의 평균 성능 지표들은 다음과 같다:

- Case I => 정확도: 0.718, 재현율: 0.742, F1 score: 0.729, ROC-AUC: 0.800

가장 성능이 우수한 상기 경우의 모델에 있어서, 2019년을 계절별로 분리하지 않고 2019년 전체 데이터 세트에 대한 예측 모델 (Case II)의 성능 지표들은 다음과 같다:

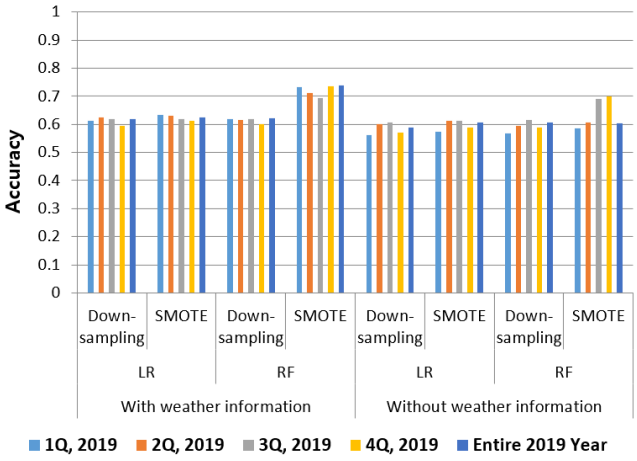
- Case II =>정확도: 0.737, 재현율: 0.652, F1 score: 0.719, ROC-AUC: 0.816

위의 성능 결과들을 비교했을 때, 분기별로 나눈 데이터 세트들 (Case I)에 대하여 항공편 지연 분류 예측 시 재현율과 F1 score값이 계절별로 나누지 않고 2019년 전체 연도에 대한 예측 모델 (Case II)의 재현율과 F1 score값보다 높음을 확인할 수 있다. 하지만, 정확도와 ROC-AUC값은 반대 현상을 보여주고 있다. 이는 계절별로 나눈 데이터 세트들에 대하여 적용한 SMOTE와 RF 기반 모델 (Case I)이 2019년 전체 데이터 세트에 적용한 SMOTE와 RF 기반 모델 (Case II)보다 실제로 지연된 항공편들에 대해서 평균적으로 올바르게 예측을 더욱 잘 하고 있음을 의미하는 것이다. 하지만, 이와는 반대로 지연되지 않은 항공편들에 대한 예측 성공율이 다소 떨어지고 있음을 의미한다. 만약, ORD 국제 공항에서 지연 항공편들에 대하여 정확하게 예측을 요구하는 시스템을 개발할 경우에는 분기별로 나눈 데이터 세트들 (Case I)을 고려하는 것이 평균적으로 향상된 성능을 보여줄 수 있다.

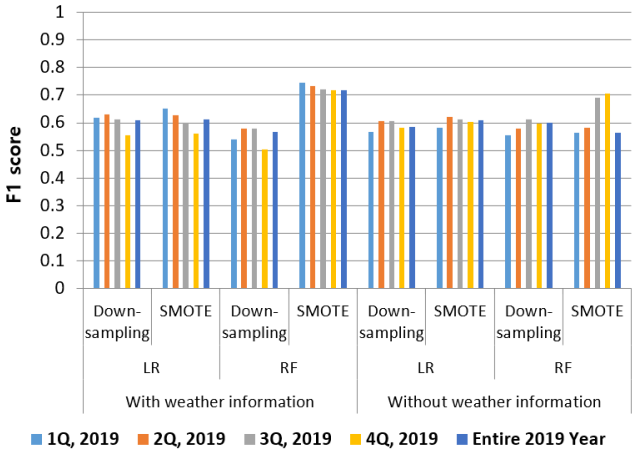
여전히 날씨 데이터가 포함된 데이터 세트에 SMOTE를 적용한 RF 기반 모델의 성능이 날씨 데이터를 고려하지 않은 어떠한 조합의 모델들보다도 성능이 우수하다. 날씨 변수의 유무에 따른 모델 성능에 대하여 직접적인 비교를 하기 위해서 SMOTE 기법이 적용된 RF 기반 모델만을 살펴보고자 한다. 기상 데이터를 포함하지 않은 SMOTE 기법을 적용한 데이터 세트에 대한 RF 기반 모델 (Case III)의 성능 지표들은 다음과 같다.

- Case III =>정확도: 0.604, 재현율: 0.501, F1 score: 0.564, ROC-AUC: 0.658

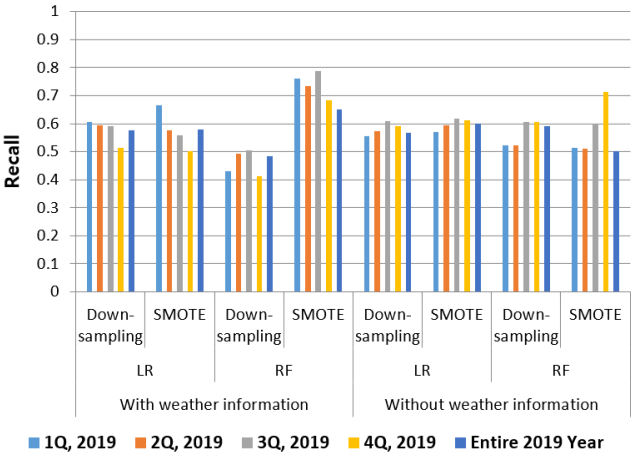
Case II와 Case III의 성능 지표 결과 분석을 통해 날씨 변수 포함 여부에 따른 항공편 지연 분류 예측 모델 성능을 비교할 수 있다. 날씨 변수를 고려한 예측 모델 (Case II)이 고려하지 않은 예측 모델 (Case III)보다 모든 성능 지표 결과값이 우수함을 확인할 수 있다.



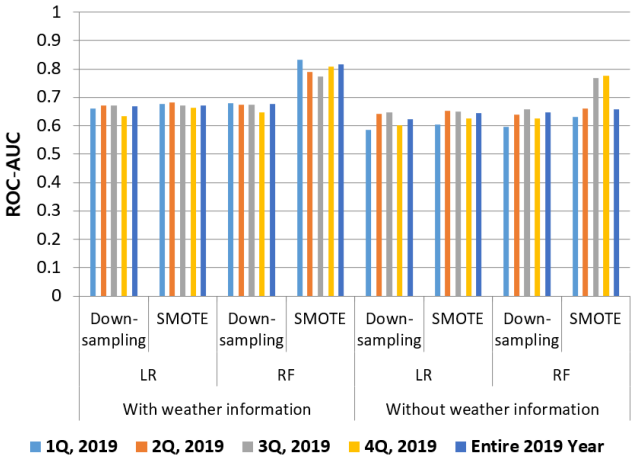
(a) Accuracy



(b) F1-score



(c) Recall



(d) ROC-AUC

그림 2. 항공편 지연 분류 예측 모델들의 성능 분석

Ⅳ. 결론

본 연구에서는 효율적인 항공편 운행을 위해 미국의 Chicago O'Hare 국제공항 (ORD)에서 출발하는 항공편 데이터 세트와 ORD 공항 주변의 날씨 데이터 세트를 이용하여 항공편 지연 분류 예측 모델을 개발하였다. 계절에 따른 영향을 확인하기 위해 분기별로 데이터 세트를 구축하여 예측 모델에 대한 분석을 진행하였다. 또한, 날씨 관련 정보가 항공편 지연

에 미치는 영향을 확인하기 위해 날씨 변수의 사용 유무에 대한 예측 모델 성능을 확인하였다. 분석 결과 oversampling의 한 방식인 Synthetic Minority Over-sampling Technique (SMOTE)를 적용한 후 Random Forest (RF) 기반 항공편 지연 분류 예측 모델 성능이 가장 우수함을 보였다. 전체적인 정확도를 고려하면, 계절별로 나누지 않고 전체 연도에 해당하는 모든 데이터를 포함한 데이터 세트에 SMOTE와 RF을 적용한 모델의 정확도가 4 계절별로 나눈 후 개별 데이터 세트들에 SMOTE와 RF을 적용한 모델들로부터 얻게 되는 평균 정확도보다 높다. 이와는 달리, 재현율과 F1 score값은 떨어졌다. 이는 분기별로 나눈 데이터 세트들에 SMOTE와 RF를 적용한 모델들이 평균적으로 분기별로 나누지 않은 전체 연도 데이터 세트에 SMOTE와 RF를 적용한 모델보다 실제로 지연된 항공편들에 대해서 올바르게 예측하는 성공율이 높다라는 것을 의미한다. 또한, 우수한 성능을 보이는 SMOTE 기법을 적용한 데이터 세트에 대한 RF 기반 모델을 살펴보았을 때, 날씨 관련 데이터들을 고려하는 것이 성능 향상을 가져옴을 볼 수 있었다.

향후 연구로써 항공편 지연 분류 예측 모델의 성능 향상을 위해서 항공편 과거 시간 데이터를 고려하고자 한다. 즉, 지연 분류를 하고자 하는 항공편의 예상 출발 시간으로부터 가까운 과거에 해당하는 항공편 데이터가 미치는 영향이 클 것이라 생각하여, 과거 시간대마다 가중치를 달리하여 항공편 지연에 대한 예측을 함으로써 성능 향상을 가져올 것이라 기대한다.

ACKNOWLEDGEMENT

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) in 2024 (2023-0-00054).

참 고 문 헌

[1] Federal Aviation Administration, “Air Traffic by the Number,” Apr. 2023, Available: https://www.faa.gov/air_traffic/by_the_numbers/media/Air_Traffic_by_the_Numbers_2023.pdf

[2] Bureau of Transportation Statistics, Available: https://www.transtats.bts.gov/airports.asp

[3] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris, “Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms,” IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Dec. 2016.

[4] Seongeun Kim and Eunil Park, “Prediction of Flight Departure Delays Caused by Weather Conditions Adopting Data-driven Approaches,” Journal of Big Data, vol. 11, no. 1, Jan. 2024.

[5] Visual Crossing, “Weather Query Builder,” Available: https://www.visualcrossing.com/weather/weather-data-services/chicago%2 0o'hare/metric/last15days

[6] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas, “Handling Imbalanced Datasets: A Review,” GESTS International Transactions on Computer Science and Engineering, vol. 30, 2006.

[7] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant, "Applied Logistic Regression,” John Wiley and Sons, Mar. 2013.

[8] Leo Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5-32, Oct. 2001.