Predicting Facebook Most-Likely Location Using Ensembled Approach in Microsoft Azure

[1]Shobhana Kashyap

[1] Assistant Profesor in Department of Computer Science & Engineering, Poornima Group of Institutions, Jaipur

[1]shobhana.kashyap@poornima.org

Abstract

With large dataset, it is very difficult to handle to work on the standalone machine. To cope with this kind of problem cloud environment is used. There is various cloud platforms for machine learning. Microsoft Azure is one of the popular platforms. In this pursuit, Facebook check-ins dataset has been used, different machine learning algorithm has been executed on the given dataset to foresee accuracy for the most famous area. Two Class Boosted Tree model has been chosen as a powerful model since it gives the most astounding accuracy of 85.8% taken after by Two Class Decision Jungle model and Two Class Bayes Point Machine model. Further, these 3 models are being gathering, which has expanded the general accuracy to 86.91%. The trial comes about have additionally been assessed utilizing 9,768 occasions that obviously approve the most extreme accuracy through Ensembling and least execution time in cloud condition.

Keywords*—Cloud Computing; Classification; Big Data; Machine Learning; Microsoft Azure.*

1. Introduction

Cloud computing is the conveyance of on demand registering assets. . With the help of computing, transfer of data over server and set a virtual environment. This give us ease of entering and storing a big amount of data in system. In this paper, it is discussed that why we are using this cloud technology, the first reason is the giant size of dataset second reason is the accuracy of the dataset. When we run our model then it listed out the results in fractions of seconds which is quicker than classical model. Cloud computing has various applications in fields of instruction, person to person communication, and pharmaceutical. In any case, the advantage of the cloud for therapeutic reasons for existing is consistent, especially due to the colossal information produced by the social insurance industry. [4][11]This enormous information can be overseen through huge information investigation, and concealed examples can be removed utilizing machine learning methodology. Specifically, the most recent issue in the therapeutic area is the forecast of heart maladies, which can be settled through the perfection of machine learning and cloud computing. Henceforth, an endeavor has been settled on to propose a decision support model how that can help medicinal specialists in foreseeing coronary illness in light of the verifiable information of patients. [21]Different machine learning calculations have been actualized on the coronary illness dataset to foresee accuracy for coronary illness.

For our research we are using facebook check-ins dataset, this dataset having 6 attributes. Attribute 6 is set as target data. This dataset is taken from Facebook for its Kaggle Recruiting Event. The dataset using for investigation of more than 100,000 spots situated in 10 by 10 km square matrix in the recreated world. The conspicuous explanation behind this dataset to be of significance is that the information utilized as a part of this dataset has been gathered more than two years and the example information speaks to the entire dataset which thus helps in building the model. Testing, as well as training information, are accessible in the .csv record with 6 qualities introduces the information. Generally, numeric esteems are utilized as information. For the analysis dataset that is utilized as a part of this examination consider 1 by 1 km square lattice. This decreases dataset and location_id.

2. LITERATURE REVIEW

In this section, we highlight the basic concentrations related to region recommendations utilizing Facebook Check Ins. [1][6][20] This early on session gives a wide review of Azure and the administrations it offers. It additionally talks about cloud computing when all is said in done and courses in which the cloud can be an advantage for analysts. At the conclusion, understudies enact their Azure Passes and investigate the Azure Entry, which is the essential tool used to oversee Azure assets.

This research study also discussed that, Azure Machine Learning is a capable tool for performing prescient examination on extensive volumes of semi-organized information. In this module, understudies utilize the intelligent Azure Machine Learning Studio to manufacture, prepare, and score a model. At that point, they set the model to work performing prescient investigation.[2] With cutting edge abilities, free get to, solid support for R, cloud facilitating advantages, simplified improvement and numerous more elements, Azure ML is prepared to take the consumerization of ML to the following level.[3] Cloud computing, that is putting forth PC resources and assets as an administration rather than an item, whereby shared assets like sound records, pictures, video documents, information, programming, and other data are given to the majority of our gadgets, be it an advanced tablet, mobile phone PC over the internet or web and is an innovation unrest giving adaptable IT utilization in a cost proficient and pay-per-utilize way. More productive methods for sharing data and teaming up can offer a genuine upper hand.

This proposition is to examine, conceptualize, incorporate and examination of two cloud-based frameworks i.e. Amazon EC2 and Microsoft Azure to store and recover documents in a cloud. In this research, the user talks about the advantages, disadvantages and furthermore make the correlation of two clouds based framework i.e. Amazon EC2 and Microsoft Azure and furthermore talk about how their cloud registering methodologies are being drawn nearer and how it impacts the eventual fate of processing.

3. OBJECTIVE

The first objective of this research is to work with a large dataset that don't compatible to analyze result on a standalone machine, and the second objective is to design the best algorithm for this dataset when we are using it in a cloud environment. For this, we are using Microsoft Azure platform, for our huge dataset, which gives us result with high accuracy and it takes a fraction of seconds to run the dataset. Ensembling technique is to be used to designing the best algorithm for our dataset.

4. METHODOLOGY

In this part of research paper, methodology used for building different classification model has been discussed. A machine learning API has been used that is built on the top of Hadoop, the accuracy of Facebook most popular places prediction model has been evaluated using Mahout. In Azure Machine learning environment the machine learning approach giving the highest accuracy has been used for the prediction model in this part. The expected input format is transformed into predicted classification by Mahout's classification method for realizing this task of predicting classification. The whole model has been deployed on Microsoft Azure (Platform as a Service), classification been selected as the given algorithm. The classification model of Facebook check-ins predicts the results to measure the accuracy and execution time.

Correlation analysis has been used for selecting most relevant predictor attributes. Fig. 1 shows the first approach in which data preprocessing steps are defined. Fig.2 describes a scale of possible correlations matrix.

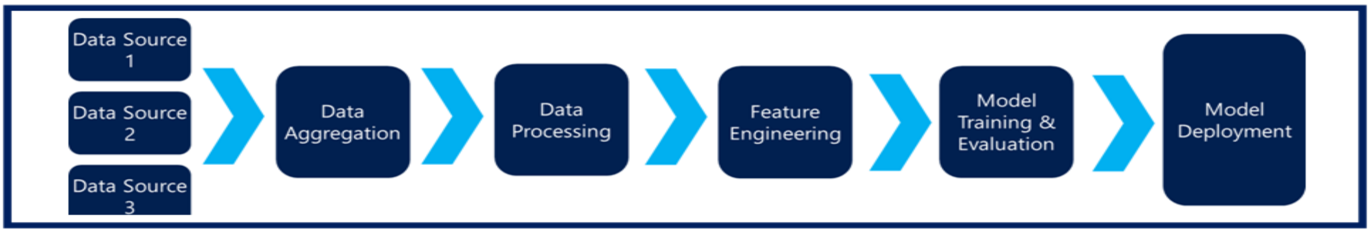
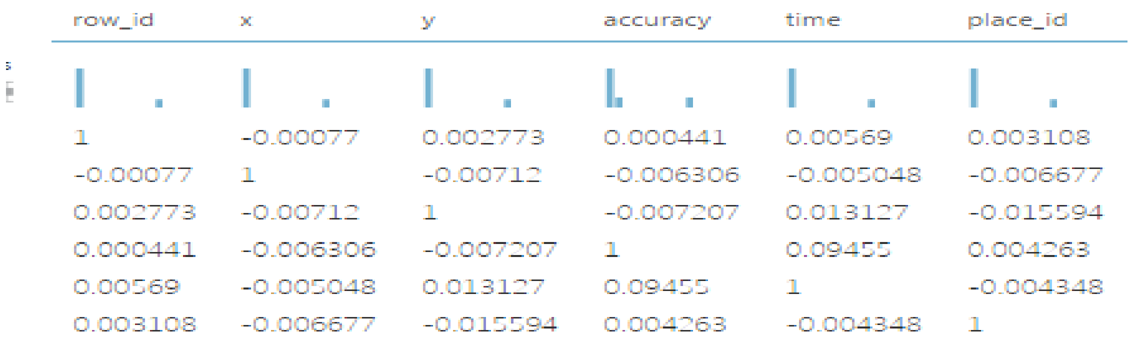


Fig. 1 Data Preprocessing steps in Microsoft azure Machine Learning Platform



row_id	x	y	accuracy	time	place_id
1	-0.00077	0.002773	0.000441	0.00569	0.003108
-0.00077	1	-0.00712	-0.006306	-0.005048	-0.006677
0.002773	-0.00712	1	-0.007207	0.013127	-0.015594
0.000441	-0.006306	-0.007207	1	0.09455	0.004263
0.00569	-0.005048	0.013127	0.09455	1	-0.004348
0.003108	-0.006677	-0.015594	0.004263	-0.004348	1

Fig. 2 Correlation between all features

5. IMPORANT RESULTS WITH DISCUSSION

5.1 Measuring performance using Azure Machine Learning Tool on Cloud Environment

Table 1 summarizes the results of confusion matrix using various machine learning algorithms. The model has been built by using 70% of data for training and 30% of data for testing. For dataset contains 9768 instances, and feature 6 set as a target data.

Various evaluation criteria like Classification Precision, Recall, F1-Square, Accuracy values have been compared and presented in a tabular form in Table 2.Fig. 3 shows the graphical representation of classification precision. Fig. 4 shows the graphical representation of classification F1-Square values. Based on classification accuracy, Fig. 5, Two Class Boosted Decision Tree algorithm gives the best accuracy of 85.8% and Two Class Neural –Network method performs the worst (83.2%) in predicting facebook most popular places

.

S.no.	Model name	TP	FN	FP	TN
1	Two –Class Support Vector Machine	381	396	143	2336
2	Two Class Neural –Network	1588	756	881	6543
3	Two Class Logistic Regression	1270	1074	491	6933
4	Two Class Locally-Deep Support Vector Machine	1202	1142	464	6960
5	Two Class Decision Jungle	1288	1056	452	6972
6	Two Class Decision Forest	1369	975	635	6789
7	Two Class Boosted Decision Tree	1532	812	571	6853
8	Two Class Bayes Point Machine	1299	1045	488	6936
9	Two Class Averaged Perceptron	1314	1030	513	6911

Table 1 Evaluation Criteria for Confusion Matrix

S.no.	Model name	PRECISION	RECALL	F1 SQUARE	ACCURACY
1	Two –Class Support Vector Machine	0.727	0.490	0.586	0.834
2	Two Class Neural –Network	0.643	0.677	0.660	0.832
3	Two Class Logistic Regression	0.721	0.542	0.619	0.840
4	Two Class Locally-Deep Support	0.721	0.513	0.600	0.836

	Vector Machine				
5	Two Class Decision Jungle	0.740	0.549	0.631	0.846
6	Two Class Decision Forest	0.683	0.584	0.630	0.835
7	Two Class Boosted Decision Tree	0.728	0.654	0.689	0.858
8	Two Class Bayes Point Machine	0.727	0.554	0.629	0.843
9	Two Class Averaged Perceptron	0.719	0.561	0.630	0.842

Table 2 Evaluation Criteria

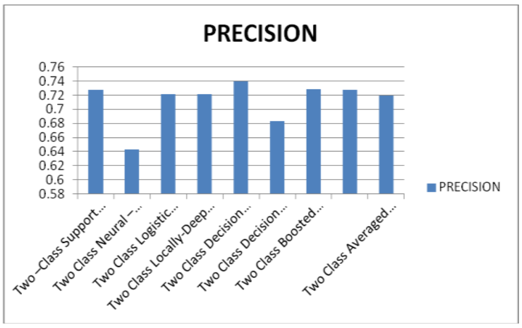


Fig. 3 Diagrammatical representation of Classification Precision

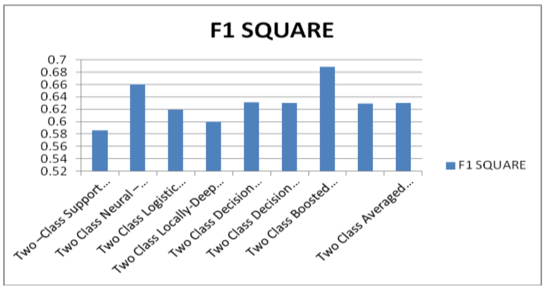


Fig. 4 Diagrammatical representation of Classification F1-Square

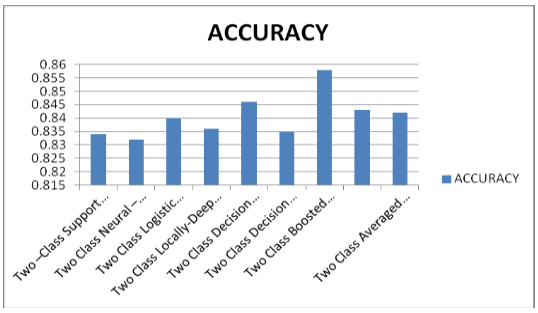


Fig. 5 Diagrammatical representation of Classification Accuracy

5.2 Measuring performance of the best algorithms using the concept of Ensembling

Ensemble technique includes consolidating various multiple predictions determined by various learning algorithms so as to make a stronger overall prediction and get best outcomes. Ensembling is one of the more popular supervised learning method of machine learning as which can be trained and additionally utilized for making predictions. Ensembling tends to yield predominant outcomes when there is a huge assorted quality among the models is so far utilized. According to the evaluation criteria are appeared in Table 2, three models have been chosen as the best models, such as Two Class Boosted Tree model, Two Class Decision Jungle model and Two Class Bayes Point Machine model. In this manner, these three models are further ensemble as to upgrade the accuracy of prediction. An ensemble model indicates most extreme classification accuracy of 87.9% which is appeared in Table 3. It likewise enhances the performance as far as Precision, Recall and F1-measure.For this we apply the method of k-fold cross validation on Ensembling result. Fig. 6 shows the confusion matrix or Ensembling model and their runs.

Runs	TP	FN	FP	TN	PRECISION	RECALL	F1 SQUARE	Accuracy
1	1723	621	856	6568	0.668	0.735	0.700	84.9
2	1518	826	553	6871	0.733	0.648	0.688	87.9
3	1848	496	1110	6314	0.625	0.788	0.697	83.6
4	1677	667	775	6649	0.684	0.715	0.699	85.2
5	1760	584	919	6505	0.657	0.751	0.701	84.6
6	1788	556	974	6450	0.647	0.763	0.700	84.3
7	1687	657	797	6627	0.679	0.720	0.699	85.1
8	1565	779	612	6812	0.719	0.668	0.692	85.8
9	1706	638	831	6593	0.672	0.728	0.699	85.0
10	1808	535	1037	6387	0.636	0.772	0.697	83.9

Table 3 Ensembling criteria using k-fold cross validation method

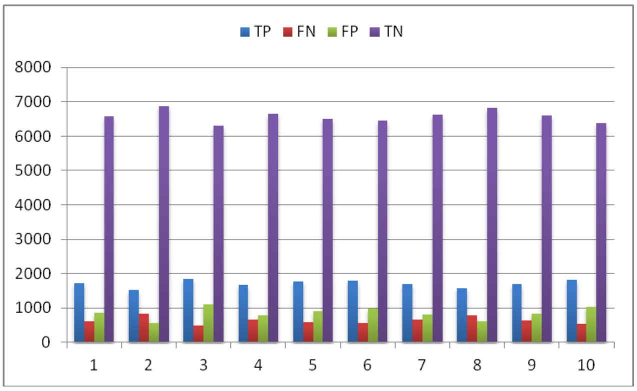


Fig. 6 Confusion Matrix

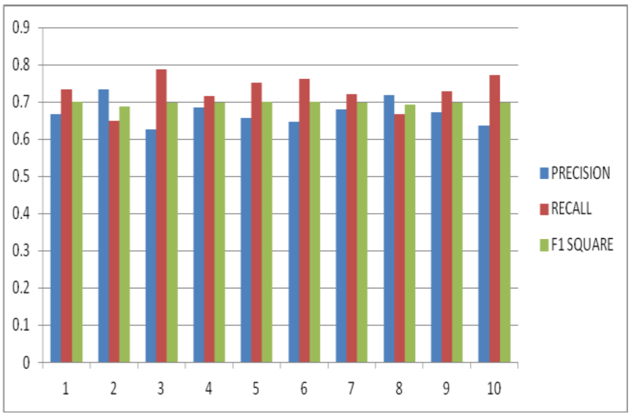


Fig. 7 Precision, Recall, and F1-Square

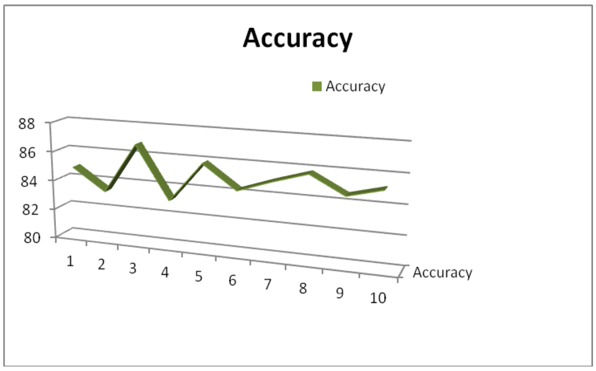


Fig. 8 Accuracy

6. CONCLUSION

In Facebook checkins dataset, the proposed model increases the prediction accuracy of target data as compared to the existing technique. In this pursuit, three models i.e. Two Class Boosted Tree model, Two Class Decision Jungle model and Two Class Bayes Point Machine model are used to create multilevel ensemble model. A novel multilevel ensemble model is developed for prediction and it produces high accuracy. It also produced the classification parameter such as Recall, Precision and F1-Square.The proposed model is compared with existing Facebook checkins model and validated on benchmark dataset. To check the robustness of proposed model, repeated k-fold cross validation is used. The proposed ensemble model achieves performance comparable to the top solutions on Kaggle, with limited feature engineering. The main take-away from this project was that careful selection of priors makes a big difference in predicting check-ins for this dataset.

7. References

[1.] Shen, Z., Jia, Q., Sela, G. E., Rainero, B., Song, W., van Renesse, R., & Weatherspoon, H. (2016, October). Follow the sun through the clouds: Application migration for geographically shifting workloads. In Proceedings of the Seventh ACM Symposium on Cloud Computing (pp. 141-154). ACM.

[2.] Luckow, A., Mantha, P., Romanus, M., & Jha, S. (2012). Pilot-Abstractions for Data-Intensive Cloud Applications.

[3.] Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1), 24.

[4.] Mazumdar, P., Agarwal, S., & Banerjee, A. (2016). Microsoft Azure Storage. In Pro SQL Server on Microsoft Azure (pp. 35-52). Apress, Berkeley, CA.

[5.] Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., ... & Haridas, J. (2011, October). Windows Azure Storage: a highly available cloud storage service with strong consistency. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (pp. 143-157). ACM.

[6.] Wilder, B. (2012). Cloud architecture patterns: using microsoft azure. " O'Reilly Media, Inc.".

[7.] Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., ... & Haridas, J. (2011, October). Windows Azure Storage: a highly available cloud storage service with strong consistency. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (pp. 143-157). ACM.