Airfare Prices Prediction Using Machine Learning Techniques

K. Tziridis, Th. Kalampokas, G.A. Papakostas
HUMAIN-Lab
Department of Computer and Informatics Engineering
Eastern Macedonia and Thrace Institute of Technology
Kavala, Greece
email: { kenaaske, theokala, gpapak}@teiemt.gr

K.I. Diamantaras
Department of Information Technology
TEI of Thessaloniki
Sindos, Greece
e-mail: kdiamant@it.teithe.gr

***Abstract*—This paper deals with the problem of airfare prices prediction. For this purpose a set of features characterizing a typical flight is decided, supposing that these features affect the price of an air ticket. The features are applied to eight state of the art machine learning (ML) models, used to predict the air tickets prices, and the performance of the models is compared to each other. Along with the prediction accuracy of each model, this paper studies the dependency of the accuracy on the feature set used to represent an airfare. For the experiments a novel dataset consisting of 1814 data flights of the Aegean Airlines for a specific international destination (from Thessaloniki to Stuttgart) is constructed and used to train each ML model. The derived experimental results reveal that the ML models are able to handle this regression problem with almost 88% accuracy, for a certain type of flight features.**

Keywords—machine learning; prediction model; airfare price; pricing models.

I. INTRODUCTION

Nowadays, the airline corporations are using complex strategies and methods to assign airfare prices [1], [2] in a dynamic fashion. These strategies are taking into account several financial, marketing, commercial and social factors closely connected with the final airfare prices.

Due to the high complexity of the pricing models applied by the airlines, it is very difficult a customer to purchase an air ticket in the lowest price, since the price changes dynamically.

For this reason, several techniques [3], [4], able to provide the right time to the buyer to purchase an air ticket by predicting the airfare price, have been proposed recently. The majority of these methods are making use of sophisticated prediction models from the computational intelligence research field known as Machine Learning (ML).

More precisely, Groves and Gini [4] applied PLS regression model to optimize airline ticket purchasing, with 75.3% accuracy (acc.). Papadakis [5] predicted if the price of the ticket will drop in the future, by handling the problem as a classification task using Ripple Down Rule Learner (74.5% acc.), Logistic Regression (69.9% acc.) and Linear SVM (69.4% acc.) ML models. Janssen [6] proposed a linear quantile mixed regression model to predict air ticket prices with acceptable performance for cheap tickets many days before departure. Ren, Yang and Yuan [7], studied the performance of Linear Regression (77.06% acc.), Naïve Bayes (73.06% acc.), Softmax Regression (76.84% acc.) and SVM (80.6% acc. for two bins) models in predicting air ticket prices.

All the aforementioned works applied only a small number of ML models, with emphasis to some classical ones, to predict the airfare prices of airlines worldwide. However, to the authors' best knowledge, the performance of the state of the art ML models to this problem is still unexplored.

The contribution of the proposed paper is summarized to the following items: (1) airfare prices prediction in Greece for the first time, (2) investigation of the features influence to the airfare prices and (3) performance analysis of the state of the art ML models in airfare prediction.

The rest of this paper is organized as follows: Section II, presents some basic information regarding machine learning, and how ML can approach the problem of airfare price prediction. Section III describes the current research from a theoretical perspective and Section IV discusses the experimental approach of the used models, as well as their results. Finally, Section V concludes the overall study and points out some research directions for future work.

II. MACHINE LEARNING

Machine Learning is one of the most hot research topics in computer science and engineering, which is applicable in many disciplines. It provides a collection of algorithms, methods and tools able to embody some kind of intelligence to machines.

The power of ML is the provided modeling tools, which are able to be trained, via a learning procedure, with a set of data describing a certain problem and to respond to similar unseen data with a common way.

Some well-known ML models are Multilayer Perceptrons (MLPs), Radial Basis Function (RBF) and Generalized Regression (GRNN) neural networks [8], Support Vector Machines (SVMs) [8], Decision Trees (DTs) [9], Extreme Learning Machine (ELMs) [10], etc.

One of the reasons that ML has attracted scientists from several disciplines is its ability to provide human-like

ISBN 978-0-9928626-7-1 © EURASIP 2017

intelligence to machines as the amount of data used during learning increases. However, the increase of the training data needs parallel implementations [11] of the ML algorithms using specialized software and/or hardware platforms.

In the context of machine learning, there are two possible alternatives for handling the problem of airfare pricing prediction. The first approach tackles the prediction of air tickets prices as a *regression* problem, while the second one transforms it to a *classification* task. The former strategy is usually applied for the prediction of the exact air ticket price, since the regression models try to approximate a function that describes the mapping law between data features and airfare prices. The later approach cannot predict the exact air ticket prices, but can provide decisions regarding the range of a price or a decision to buy or not the ticket with the specific price.

In this paper, the first case of airfare price prediction via regression is considered, since little attention has been paid in evaluating the state of the art regression ML models for that problem.

III. Current Study

Initially, the Greek Aegean Airlines [12] company and its flight, from Thessaloniki to Stuttgart, is selected as the case study of our investigation.

The current study consists of four distinctive phases: (1) the selection of the flight features that influence the airfare prices, (2) the collection of enough flights data which will be used to train and test the applied ML models, (3) selection of the regression ML models being compared and (4) experimental evaluation of the ML models.

Each processing phase is discussed in more detail in the following:

Phase 1 (Feature Selection) - During this phase the most informative features of a flight that determine the prices of the air tickets are decided. This phase is very important since it defines the problem under solving.

For every flight the following features were considered:

- F1: Feature 1 - departure time.
- F2: Feature 2 - arrival time.
- F3: Feature 3 - number of free luggage (0, 1 or 2).
- F4: Feature 4 - days left until departure.
- F5: Feature 5 - number of intermediate stops.
- F6: Feature 6 - holiday day (yes or no).
- F7: Feature 7 - overnight flight (yes or no).
- F8: Feature 8 - day of week.

It is worth to note that the influence of some critical features from the above list will be examined through an "one-leave-out" rule. We also like to clarify that the feature F4 indicates the number of days between the ticket purchase and the day of the flight.

Phase 2 (Data Collection) - In this study, our interest is focused on the prediction of a single airfare price without return. For the sake of the experiments a set of flights to the same destination (from Thessaloniki to Stuttgart) for the period between December and July, is collected. For each flight the eight features (F1:F8) were manually collected from the Web, 1814 flights were recorded totally and are available in [13].

Phase 3 (ML Models Selection) - Eight state of the art regression ML models [8], [10], [14], [15], [16] were selected for the current study and applied to the same data of flights. The ML models compared in this work are the following:

- Multilayer Perceptron (MLP).
- Generalized Regression Neural Network.
- Extreme Learning Machine (ELM).
- Random Forest Regression Tree.
- Regression Tree.
- Bagging Regression Tree.
- Regression SVM (Polynomial and Linear).
- Linear Regression (LR).

Phase 4 (Evaluation) - The 1814 flights collected in phase 2, were used in a 10-fold cross-validation procedure to train the aforementioned ML models. The performance indices used to compare the models are the prediction accuracy (% - MSE between the desired and predicted prices) and the time in seconds, needed to train each model.

IV. Simulations

For the sake of the experiments a set of simulations were arranged and executed under the MATLAB environment in a i5-750 2.67 GHz PC with 8GB memory. The configuration of the ML models was decided by applying grid search and is summarized in Table I.

TABLE I. Models Configuration

ML Model	***Configuration***
Multilayer Perceptron (MLP)	3 hidden layers 5 nodes each layer
Generalized Regression Neural Network	spread=1.0
Extreme Learning Machine	10 neurons
Random Forest Regression Tree	300 weak classifiers (decision trees)
Regression Tree	MinParentSize=10 MinLeafSize=3 MaxNumSplits=45
Bagging Regression Tree	500 weak classifiers (decision trees)
Regression SVM (Polynomial)	order=3
Regression SVM (Linear)	stochastic gradient descent solver
Linear Regression	dual stochastic gradient descent solver

ISBN 978-0-9928626-7-1 © EURASIP 2017

A 10-fold cross-validation procedure was applied to all the experiments and the mean performance of each model is presented in this section.

The performance of all models for the case of the entire feature set (eight features) is presented in Table II, with the highest performed model being bold faced.

TABLE II. RESULTS WITH ALL FEATURES

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	80.28	20.88
Generalized Regression Neural Network	66.83	0.13
Extreme Learning Machine	68.68	0.05
Random Forest Regression Tree	85.91	5.50
Regression Tree	84.13	0.04
Bagging Regression Tree	**87.42**	17.05
Regression SVM (Polynomial)	77.00	1.23
Regression SVM (Linear)	49.40	0.34
Linear Regression	57.25	0.10

From the results of Table II it is obvious that the "Bagging Regression Tree" model outperforms the other models, while its training is quite fast. Moreover, the "Random Forest Regression Tree" seems to be an alternative choice since it shows similar performance in less time.

In order to analyze the influence of the used features to the prediction accuracy of the models, the same experiment is repeated several times by leaving out some features, one at a time. In this context the first two time features were removed and the experiment is repeated with six features (F3:F8). The corresponding results are presented in Table III.

TABLE III. RESULTS WITHOUT F1 & F2 FEATURES

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	75.49	16.31
Generalized Regression Neural Network	66.25	0.14
Extreme Learning Machine	67.18	0.05
Random Forest Regression Tree	**79.49**	10.6
Regression Tree	78.76	0.06
Bagging Regression Tree	77.50	15.07
Regression SVM (Polynomial)	78.12	0.87
Regression SVM (Linear)	44.95	0.42
Linear Regression	57.19	0.23

From the above results it is obvious that almost all models shown lower (up to 10%) prediction accuracy and greater execution time. These results reveal that the timing features "departure time" and "arrival time" influence significantly the airfare prices. Furthermore, the increase of the execution time means that the training procedure converges quite later for almost all the models.

Table IV summarizes the performance of the models when the "day of week" feature (F8) is omitted during training.

TABLE IV. RESULTS WITHOUT F8 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	81.58	5.65
Generalized Regression Neural Network	66.83	0.32
Extreme Learning Machine	66.88	0.086
Random Forest Regression Tree	86.18	5.28
Regression Tree	84.22	0.02
Bagging Regression Tree	**87.59**	13.73
Regression SVM (Polynomial)	79.38	0.98
Regression SVM (Linear)	60.64	0.02
Linear Regression	57.07	0.05

In this case, we observe that all models were not affected as much as previously, except "Regression SVM" with Linear kernel. Therefore, one can conclude that the "day of week" does not influence airfare prices.

Table V, presents the performance of the models without using the "overnight flight" feature (F7). The outcomes of this experiment reveal that this feature is not related with the price of the air ticket, since the models perform similarly or even worse with the case of using all features. Only the "Multilayer Perceptron" and the "Regression SVM" with Linear Kernel seems to be affected significantly by this feature.

TABLE V. RESULTS WITHOUT F7 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	72.8	5.98
Generalized Regression Neural Network	66.14	0.34
Extreme Learning Machine	64.88	0.06
Random Forest Regression Tree	86.15	6.15
Regression Tree	84.22	0.059
Bagging Regression Tree	**87.93**	15.34
Regression SVM (Polynomial)	77.91	0.17
Regression SVM (Linear)	57.69	0.06
Linear Regression	57.92	0.02

Next, we are leaving out the "holiday day" feature (F6), and the models are executed again. Their performance is

ISBN 978-0-9928626-7-1 © EURASIP 2017

similar with that of the first experiment, as illustrated in Table VI.

TABLE VI. RESULTS WITHOUT F6 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	77.94	5.74
Generalized Regression Neural Network	66.31	0.25
Extreme Learning Machine	68.5	0.05
Random Forest Regression Tree	86.17	5.54
Regression Tree	84.13	0.02
Bagging Regression Tree	**87.60**	16.47
Regression SVM (Polynomial)	67.2	0.15
Regression SVM (Linear)	57.69	0.05
Linear Regression	57.92	0.02

The "Bagging Regression Tree" outperforms all the models not only in this experiment, but also all the models under different feature sets examined previously. The reminder models seem not to be affected by the exclusion of "holiday day" feature.

The last experiment is executed without using the "number of intermediate stops" feature (F5), with similar results with the first experiment.

TABLE VII. RESULTS WITHOUT F5 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	78.62	7.43
Generalized Regression Neural Network	65.24	0.32
Extreme Learning Machine	66.83	0.03
Random Forest Regression Tree	86.04	4.79
Regression Tree	83.88	0.01
Bagging Regression Tree	**87.91**	16.32
Regression SVM (Polynomial)	77	0.14
Regression SVM (Linear)	49.4	0.05
Linear Regression	57.25	0.02

Concluding the previous study, one can claim that "Bagging Regression Tree", "Random Forest Regression Tree", "Regression Tree" and MLP models are the most stable models according to their accuracy scores. In addition, as far as the execution time is concerned the best models are "Random Forest Regression Tree" and "Regression tree".

V. CONCLUSION

This paper reported on a preliminary study in "airfare prices prediction". We gathered airfare data from a specific Greek airline corporation (Aegean Airlines) from the web and showed that it is feasible to predict prices for flights based on historical fare data. The experimental results show that ML models are a satisfactory tool for predicting airfare prices. Other important factors in airfare prediction are the data collection and feature selection from which we drew some useful conclusions. From the experiments we concluded which features influence the airfare prediction at most.

Apart from the features selected, there are other features that could improve the prediction accuracy. In the future, this work could be extended to predict the airfare prices for the entire flight map of the airline. Additional experiments on larger airfare data sets are essential, but this initial pilot study highlights the potential of Machine Learning models to guide consumers to make an airfare purchase in the best market period.

REFERENCES

[1] P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203, 2009.

[2] P. Malighetti, S. Paleari and R. Redondi, "Has Ryanair's pricing strategy changed over time? An empirical analysis of its 2006–2007 flights," Tourism Management, vol. 31, no. 1, pp. 36-44, 2010.

[3] W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.

[4] W. Groves and M. Gini, "An agent for optimizing airline ticket purchasing," 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 06 - 10, 2013, pp. 1341-1342.

[5] M. Papadakis, "Predicting Airfare Prices," 2014.

[6] T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.

[7] R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price," Technical Report, Stanford Univerisy, 2015.

[8] S. Haykin, Neural Networks – A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.

[9] S.B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.

[10] G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1-3, pp. 489-501, 2009.

[11] G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm," Journal of Parallel and Distributed Computing, vol. 99, pp. 90-99, 2017.

[12] Aegean Airlines, https://en.aegeanair.com.

[13] https://github.com/humain-lab/airfare_prediction.

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[15] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Boca Raton, FL: CRC Press, 1984.

[16] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines," Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.

ISBN 978-0-9928626-7-1 © EURASIP 2017