Faculty of Technology and Society
Department of Computer Science and Media Technology

Master Thesis Project 15p, Spring 2020

Efficient flight schedules with utilizing Machine Learning prediction algorithms

By
Mashhood Vandehzad

Supervisor:
Reza Malekian

Examiner:
Johan Holmgren

Contact information

Author:
Mashhood Vandehzad
E-mail: vandehzad@gmail.com

Supervisor:
Reza Malekian
E-mail: reza.malekian@mau.se
Malmo University, Departament of Computer Science

Examiner:
Johan Holmgren
E-mail: johan.holmgren@mau.se
Malmo University, Departament of Computer Science

Abstract

While data is becoming more and more pervasive and ubiquitous in today's life, businesses in modern societies prefer to take advantage of using data, in particular Big Data, in their decision-making and analytical processes to increase their product efficiency. Software applications which are being utilized in the airline industry are one of the most complex and sophisticated ones for which conducting of data analyzing techniques can make many decision making processes easier and faster. Flight delays are one of the most important areas under investigation in this area because they cause a lot of overhead costs to the airline companies on one hand and airports on the other hand. The aim of this study project is to utilize different machine learning algorithms on real world data to be able to predict flight delays for all causes like weather, passenger delays, maintenance, airport congestion etc in order to create more efficient flight schedules. We will use python as the programming language to create an artifact for our prediction purposes. We will analyse different algorithms from the accuracy perspective and propose a combined method in order to optimize our prediction results.

Acknowledgement

The fulfillment of this research study would not have been possible without the constant guidance and help of certain people.

First and foremost I wish to express sincerest gratitude to my supervisor Reza Malekian for the constant guidance, invaluable support and also for time spent in this dissertation.

Secondly, I wish to express most sincere appreciation to my examiner Johan Holmgren for the comments received which without them, I could not have completed this project.

I would also like to thank Nils Genell, Paria Aghaeifar and Niklas Nordin from Aviolinx company for the opportunity of creating this research study with them.

Last but certainly not the least, I would like to appreciate my parents, my family and friends for their unconditional support and also for inspiring and motivating me through my path.

Contents

List of Figures

List of Acronyms

AI	Artificial Intelligence
CM	Combined Method
GFR	Grand Father Rights
LR	Linear Regression
ML	Machine Learning
PA	Prediction Accuracy
PE	Prediction Error
RBF	Radial Basis Function
SVM	Support Vector Machine
SVR	Support Vector Regression

Chapter 1

Introduction

We are living in a world where massive amounts of data get generated everyday and collections of data are growing exponentially. The process of converting raw data to meaningful and utilizable information can be a sophisticated process in which technology is playing a very important role today. Nowadays we can implement machine learning techniques on different types of data for the purpose of creating new knowledge from data sources that where not being used before.

Since we as humans want to give the learning ability of human brain to machines and also these new techniques are becoming an inseparable part of latest information systems developed world wide; Therefore, the implementation of machine learning techniques are one of the most preponderantly under research areas these days, especially in software-intensive companies. One of the areas in which machine learning is mostly being used is for prediction purposes. Prediction of future events can help humans prepare themselves to react more effectively and efficiently and makes the decision making process more facile. Examples of software systems that prediction methods could be favorable for is stock market predicting or airline industry flight scheduling systems.

This thesis project has been made in collaboration with a company called Aviolinx in Malmo Sweden. Their line of work is software manufacturing for airline industry. Currently they have not utilized any machine learning tech-

niques or algorithms in their software slot scheduling system so we decided to build a small artifact that would be capable of predicting future flight delays utilizing four different machine learning algorithms and real world flight delays data. We thought with the utilization of this artifact, airport coordinators and airline companies together would be able to deploy future flight schedules in a more efficient manner. We will have a comparison chart at the end of this thesis project that analyzes the strength of each of the four ML algorithms implemented in our artifact. Another perspective considered for this thesis project is that the simultaneous comparison of these well-known prediction algorithms can contribute to the research community for future and further investigations.

1.1 Research Questions

- RQ1: How machine learning algorithms can be leveraged on flight delays data for prediction of flight scheduling systems?
- RQ2: How precise would different algorithms be able to predict comparing to the real flight delays?

1.2 Goals

The aim of this project is to study how effectively machine learning techniques can contribute to airline industry scheduling systems by forecasting flight delays. Due to the problems airport coordinators encounter everyday with airline companies regarding flight delays and the continues need for changing and revising flight gantt schedules. Therefore, we propose to develop an artifact utilizing machine learning in order to predict future delays. This artifact may lead to mitigation of flight delays' related schedule revising and consequently reduce costs. Since we will evaluate four different algorithms, we will analyse and compare the preciseness of each of the implemented algorithms by utilizing graphs, charts and tables. So the reader

of this research study would realize which algorithms are most suitable to be utilized in flight delay prediction systems.

1.3 Expected Results

First we expect that after finishing with this thesis project, the literature review and the results of developing our artifact, we would have a clear understanding of how accurate any of the chosen ML algorithms can predict flight delays and show the comparison amongst them utilizing graphs and tables. Afterwards we expect that this project could help airline software manufacturers have better insights into implementation of machine learning prediction techniques into their systems. Machine learning can contribute them to use raw data and convert it to delay prediction information that may help them in many different ways in better and more precise scheduling of their future flights.The more efficient scheduling systems they make the more financially beneficial the system becomes.
As discussed with Aviolinx company, This prototype can be utilized as a stand alone reporting system or integrated into their slot scheduling software with further developments as a new feature for airline companies to use.

1.4 Motivation

The most important motivation for this project is to help airline industry software manufacturers utilizing machine learning [1] algorithms inside their software products enabling them for prediction capabilities, so their customers which are airline companies would be able to have near future forecasts and use this knowledge to prevent delays by different causes in their schedules. Delays can impose variety of costs to airline companies which in fact by reducing small amounts of delays, huge amounts of company revenue would be affected.

1.5 Outline

This master thesis consists of six chapters. In the first chapter we introduce this thesis project, the process of completing it and the goals and motivations for creating this research project in general. In the second chapter we explain details about the research methodology chosen for this thesis and how each stage of the process of this study is taken and we elaborate our research questions and the goals for making the artifact and also the results which we expect to achieve with this project. In the third chapter we will have a literature review in which we divide our investigation into two main subsections. In the first subsection we elaborate machine learning and all the algorithms related to this thesis project and in the second subsection we investigate some of the related terms and conditions of the aviation and airline industry. In the forth chapter we elaborate the design and implementation of our application and explain how it performs via graphs so the audience of this research study can visualize our methods and results. In the fifth chapter we will rigorously analyse our results and demonstrate different comparisons of the results for this study paper. Finally in the sixth chapter we discuss the conclusions and the future work that can be done to optimize and improve this project.

Chapter 2

Research Methodology

2.1 Research Method

In this chapter we describe the research methodology that has been applied to collect necessary information and to perform the research study in order to answer this master thesis's research questions. We have found Design Science [2] most suitable research methodology for this study paper and we have planned different stages of our research process as shown bellow in figure 2.1 that will be explained later in this chapter.

2.1.1 Literature Review

We will use some of the famous digital libraries such as IEEE or ScienceDirect in order to create a viable literature review and further illustrate the concepts and Machine Learning algorithms needed for this research study. In the first phase of this thesis project we will have a literature review in which we will divide our investigation areas into two different parts. In the first part we will investigate previous academic work like [3] which explain data utilization and try to evaluate different ML algorithms utilized in our project as explained in [4]. In the second part we will explain some of the important terms and conditions that exists in the airline industry like mentioned in [5], and needs to be explained to the reader of this paper in order to clarify the environment in which this artifact can be utilized. Also a brief statistics from previous

Figure 2.1: Research Method

academic work on how flight delays could affect airline industry revenues and costs.

2.1.2 Design and Creation

Design Science [2] is a research methodology used for the creation of IT prototypes regarding technology and bringing new solutions to answer organizational real world problems. These solutions could be in a form of constructs, models, methods or instantiation [2] which may later considering more assessments and analysing requirements become software, IT products or in general innovations that make use of information systems more effective. The research methodology chosen for this study is Design Science [2] because we will develop a prototype capable of analyzing real world flight

delays data and implementing machine learning algorithms to predict future flight delays. The final results of our artifact will be presented in the form of models. We will plot these models into graphs in order to visualize the investigated data sets and predicted results. We will have a rigorous testing of our artifact's results afterwards.

2.1.3 Evaluation

As proposed by Hevner [2] in an information systems research study "The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods" and that the accuracy of an IT-Artifact can be evaluated as the quality attribute. Since design is an iterative and progressive process, feedback from the evaluation phase plays a key role for the final artifact to be satisfactory for the problems it was meant to solve [2].
Therefore after finishing with the development of our prototype we need to test it rigorously to be able to analyze the prediction results. In this phase of our project we will feed the available flight delays data to our application and compare the outcomes of each algorithm by showing their strength in future prediction using bar charts with percentage of accuracy. Considering that we have some limitations accessing the real flight delays data from Aviolinx we decided to assess this artifact with the limited amount of data which is related to two different airline companies and show the reader of this paper the results in a smooth and coherent way.

2.1.4 Plotting the Models

Scholarly data is a fundamental part of a scientific research that needs to be presented in a way that reveals hidden patterns in the data to make it more analysable and understandable. Various data visualization techniques nowadays are being used in order for the readers to "create a visual expression instead of numerical complex scientific concepts or results" [6].
For visualization purposes in this research study we need to use tables, graphs, charts and different colors as explained earlier in this chapter in

order to demonstrate our models and results to make it more comprehensible for anyone who concerns about investigating how to use machine learning for prediction purposes.

Chapter 3

Literature Review

3.1 Search Keywords

Airport slots, airport scheduling software, aviation grandfather rights, slot messages, scheduling airport slots,machine learning for prediction,machine learning approaches.

In the first part of our literature review we will explain data, machine learning and the algorithms utilized in our research study.

3.2 Data

Data sets grow exponentially in size everyday because of many different methods which are now available to collect different types of data. Mobile devices, aerial sensory, software logs, cameras, microphones and many more ways that we use gathering data has resulted in the collection of massive amounts of data. "There are 2.5 quintillion bytes of data created every day, and this number keeps increasing exponentially. The world's technological capacity to store information has roughly doubled about every 3 years since the 1980s" [7]. Large amounts of data in different environments like for instance financial or medical areas, are created at high costs and unfortunately deleted because of lack of required technologies to store them. These are

valuable data which could be utilized during the production of new features for new software. This has become one of the biggest and most costly challenges for companies to come up with new solutions for storing huge amounts of usable data. However, with the advancements of technology now we can use new architectures and mechanisms to store and access valuable data for software optimization purposes [7].

3.3 Machine Learning (ML)

Machine Learning is one of the most under investigation research areas these days. Since we as humans want to make machines as intelligent as human brain which is perceived as Artificial Intelligence (AI), the concept of giving the learning capability to machines coming from the learning ability of human brain [8] is now in process more and more everyday. Machine Learning is an integral part of the AI that we use to design different kinds of algorithms in order to utilize in a variety of fields like bioinformatics, intrusion detection, Information retrieval, game playing, future predictions, marketing, malware detection and to investigate data trends and historical data relationships [9].

All the existing machine learning algorithms as shown in figure 3.1 [10] are derived from two main strategies called supervised and unsupervised. Supervised strategy is utilized when the training set comprises the data and the authentic output of the process that uses that data. An example can be when a set of problems and their solutions are given to a student in order to solve future problems alike in that area. However, the unsupervised strategy is used when the training set comprises the data but it does not contain solutions for it and the computer must resolve the problem by itself. An example can be when a set of patterns are given to a student and asking them to reveal the underlying relations that generated those patterns.

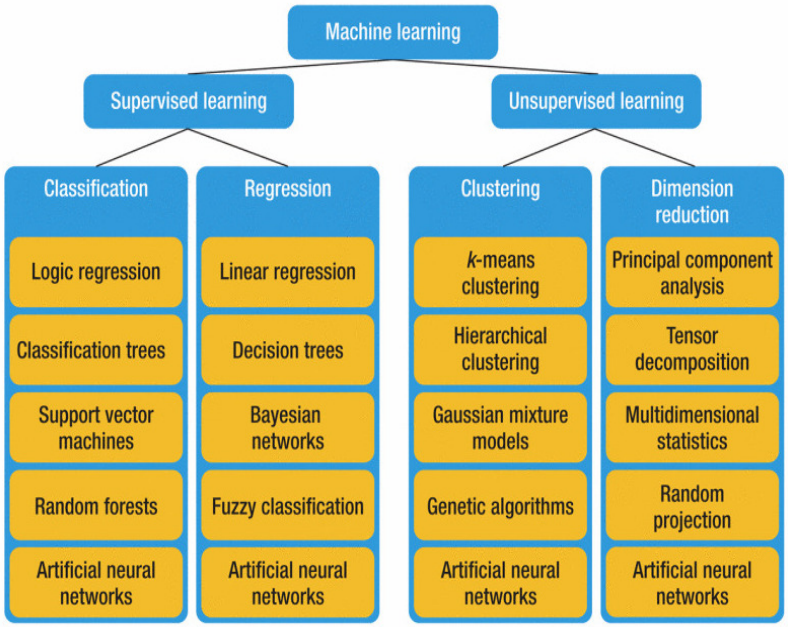

Figure 3.1: Machine-learning approaches

3.3.1 Linear Regression (LR)

Regression is an approach under the category of supervised learning. This approach can be used for prediction purposes when we want to model continues variables and make future predictions. Examples of applications using regression for prediction purposes are real-estate price prediction, stock movements forecasting, student scores predictions. In regression we have the labeled dataset that the output values are figured out utilizing the input values. The linear regression algorithm is the simplest form of regression that we try to fit a straight line (straight hyperplane) to the dataset [11].

3.3.2 Support Vector Machine (SVM)

Support Vector Machines can resolve both regression and classification problems. In this method hyperplane needs to be specified the decision boundary. The objects in this method can be separated utilizing complex mathematical

functions called kernels [11]. When we want to utilize SVM for classification purposes it is called SVC and when we want to use it for regression purposes it is called SVR [12]. In our prototype we have utilized three different types of linear and non linear SVR kernels. SVR_RBF Which is a nonlinear kernel. SVR_Lin which is a linear kernel. SVR_Poly which is a nonlinear polynomial kernel.

3.3.3 Reason for Choosing the Algorithms

For more accurate investigation of this study we needed robust algorithms capable of predicting flight delays with high accuracy. Also in order to have a balance for our comparison purposes we decided to use two non-linear algorithms which are SVR_RBF and SVR_Poly and two linear algorithms which are SVR_Lin and LR. All these algorithms are well-known for their robustness in prediction purposes [12].

In the second part of our literature review we will explain some of the terms and conditions used in the airline industry.

3.4 Airport Slots

Slots are the time intervals in which a flight will be allowed to use an airport infrastructure for landing and take-off according to a pre-coordinated schedule [5]. As an example we can say the 10 a.m. to 10.30 a.m. slot for Mondays and Thursdays is reserved for KLM airline company in Copenhagen airport for summer season 2019.

3.5 Slot Message

For the purpose of coordination between an airline company and an airport, special type of messaging system exists in which first the airline company requests a slot from the airport and then the airport coordinator approves or denies the request via messaging back. This process continues till both sides

agree on one slot for one season. An example of a slot message is shown in figure 3.2

Example

```
SCR
/AF1506
W03
15JUN
CPH
CAF802 AF810 26OCT27MAR 1234567 290AB3 FCONCE0910 1030LHRMAN JJ
RAF802 AF810 26OCT27MAR 1234567 290AB3 FCONCE0920 1050LHRMAN JJ
```

Figure 3.2: Example of a Slot Message

3.6 Airport Slot Allocation

Grandfather rights or "use-it-or-lose-it" rule means an airline company must use 80% of the slots allocated to them by an airport through one season, otherwise they will lose that slot and the slot will be released for other companies to take [5]. This means losing potentially huge amounts of money that the airline company has invested to take a specific slot, specially when it comes to more congested airports it becomes incredibly arduous to take one slot for a season.

3.7 Statistics

A study in 2010 shows that the total cost for the US airline industry in 2007 related to flight delays was $32.9 billion. This amount is consisted of $8.3 billion related to increasing of the expenses for the crew, fuel and maintenance, $16.7 billion is related to passenger time lost, $3.9 billion is related to demand loss for the passengers as the result of delays. And $4 billion is related to indirect effects of the flight delays on US economy as reduction of US GDP. These statistics illustrate the significance of the flight delays problems and the need for more efficient investigations and methods for the purpose of reducing flight delays [13].

Chapter 4

Method

The prototype developed for this research study predicts flight delays by utilizing Machine Learning. With utilizing this artifact we will be able to convert raw flight delay data into usable information by machine learning techniques, that can add more efficiency to a flight scheduling software system and this way we answer our RQ1 on how to use flight delays data. With this method, as we have discussed with Aviolinx staff, they have the option to use this artifact in two different ways. They can either integrate it into their software in the future or they can have this as a stand alone reporting system. Either way they can implement these predictions into slot messages in which airport coordinator and airline company can have an idea about upcoming month in order to schedule in a more accurate state.

4.1 Description

This prototype is written in python programming language [14]. It uses different python libraries to be able to predict delays based on four different algorithms and then plots the models into graphs for visualization purposes. We implemented our code into Google Colaboratory (also known as Colab) which is a cloud service based on Jupyter IDE giving us the advantages of investigating the purpose of our study without concerning about required configurations, GPU access and sharing our code [15].

4.2 Datasets

We have received our real world flight delay datasets from Aviolinx for which we experienced many limitations considering their regulations and also this process consumed a lot of time. The two datasets that we use in our project are related to two different airline companies which are using Aviolinx software product. After requesting the real world datasets from Aviolinx we ultimately were authorized to use 6 years flight delay data from two airline companies from 2014 to 2019 anonymously and we did accordingly. Therefore, since the data is real and we are not allowed to mention the names of these companies, we will use "Company A" and "Company B" instead of their names. Our datasets represent aggregated flight delay data by all causes for each month since January 2014 which was the limit for us to access the data. So we have access to all the data from January 2014 to December 2019 which we will use to evaluate our artifact. As an example if the data shows 12.5 minutes for December 2016 it shows that the aggregated number of minutes by all causes such as passenger delays, aircraft maintenance, airport congestion, weather etc. is 12.5 minutes for that specific airline company in December 2016.

4.3 Libraries

The libraries utilized in our code are pandas [16], numpy [17], SKlearn [4], Matplotlib [18] and Google colab [15]. We implemented our machine learning algorithms in the code utilizing three different Support Vector Regression (SVR) Methods and Linear Regression from SKlearn library in order to train our dataset and predict future delays. The main reason for using four algorithms was to be able to compare the results. In order to load our dataset we utilized Google colab library [15]. Finally when the models are set we utilized Matplotlib library [18] to plot the models into graphs.

4.4 Algorithms

In this artifact We have utilized three types of support vector regression linear and non-linear kernels and a linear regression algorithm on our data for our accuracy comparison purposes. The algorithms are as follows:

- Support Vector Regression - RBF (Radial Basis Function), which in this report we write it as SVR_RBF [4]
- Support Vector Regression - Linear, which in this report we write it as SVR_Lin [4]
- Support Vector Regression - Polynomial, which in this report we write it as SVR_Poly [4]
- Linear Regression [19]

After we load the dataset in our application we create the models and train the models utilizing the flight delays data with previously mentioned algorithms. After the models are trained we plot them into a graph and show the results. For the purpose of prediction we load all the data from the first month of 2014 to predict flight delays of the year 2019 so for example in order to predict the first month of 2019 we investigate the data for 60 month meaning that we have 60 data points.

4.5 Models

4.5.1 Model related to Company A

The first model in figure 4.1 shows the prediction results by different algorithms compared to the real value from the dataset provided from company A. This model is an example of the output of our application which shows the prediction results of delays for September 2019. The prediction results for different algorithms are shown by black rectangles at the bottom of this figure and the real value for this month is shown in the black square. As

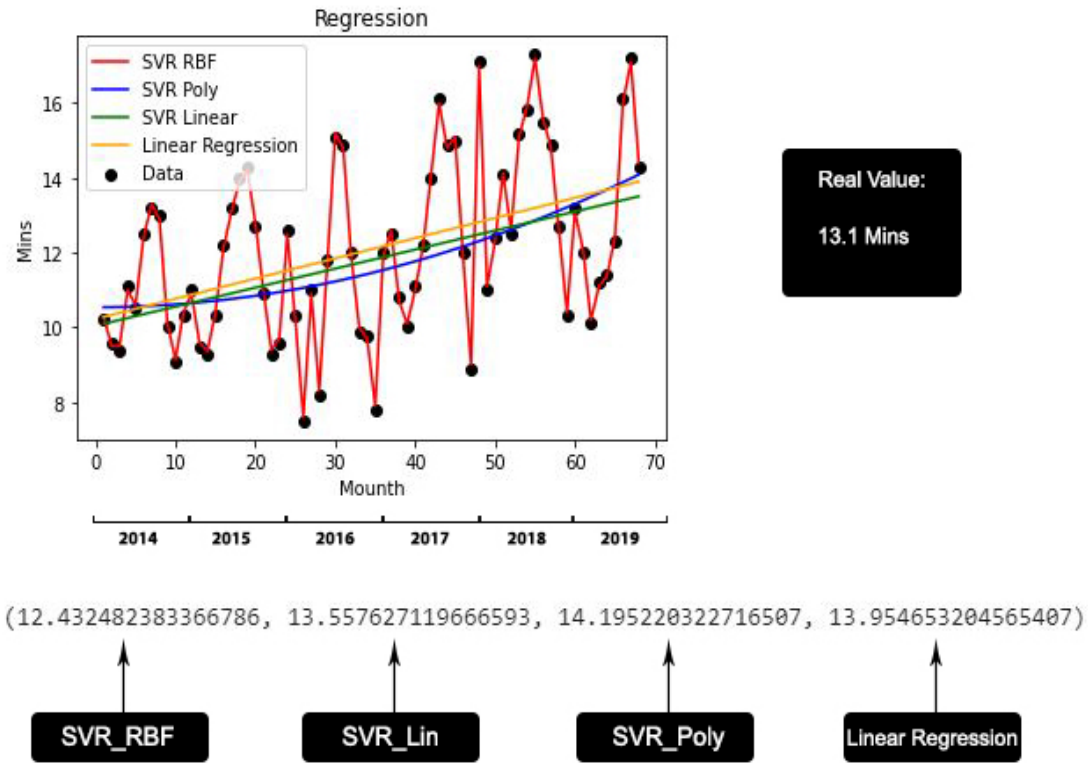

Figure 4.1: Predictions of September 2019 for Company A

illustrated in figure 4.1 our models are trained with all the data from January 2014 till August 2019 and predict September 2019 for company A. The prediction result with SVR_RBF is estimated 12.43 minutes, with SVR_Lin is estimated 13.55 minutes, with SVR_Poly is estimated 14.19 minutes and with Linear Regression is estimated 13.95 and the real value is 13.1 minutes.

4.5.2 Model related to Company B

The second model in figure 4.2 shows the prediction results by different algorithms compared to the real value from the dataset provided from company

B. This model is an example of the output of our application which shows the prediction results of delays for March 2019. The prediction results for different algorithms are shown by black rectangles at the bottom of this figure and the real value for this month is shown in the black square. As illustrated in figure 4.1 our models are trained with all the data from January 2014 till February 2019 and predict March 2019 for company B. The prediction result with SVR_RBF is estimated 13.05 minutes, with SVR_Lin is estimated 12.75 minutes, with SVR_Poly is estimated 13.25 minutes and with Linear Regression is estimated 12.76 and the real value is 12.9 minutes.

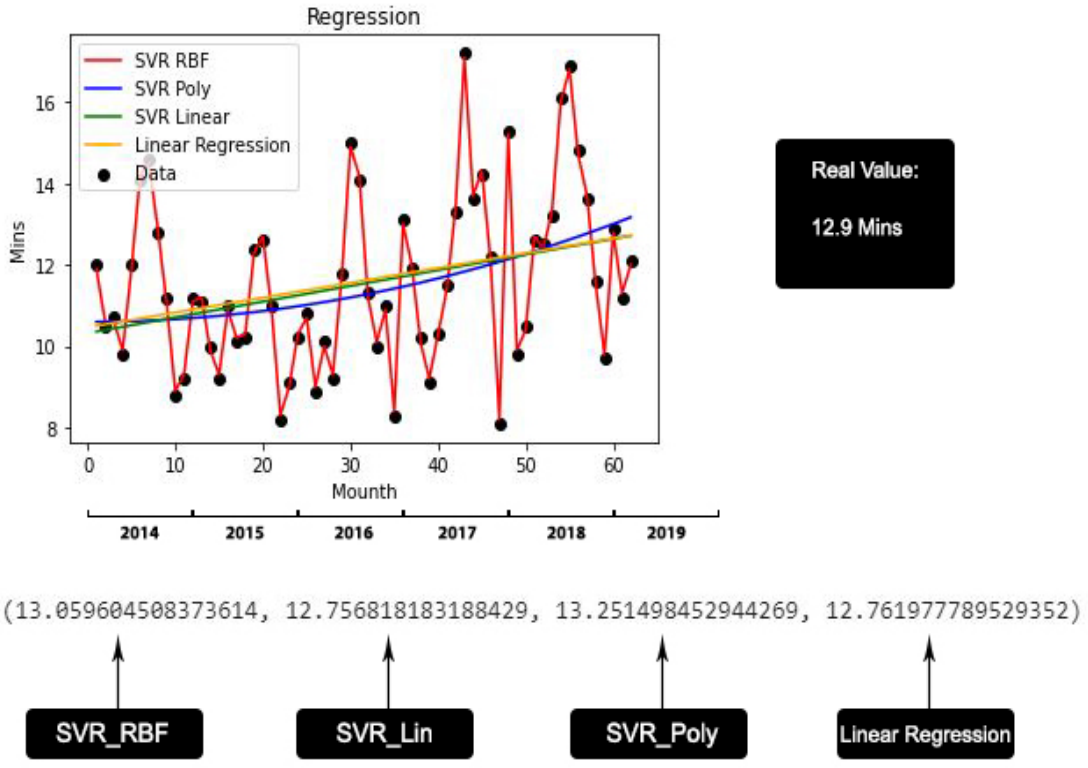

Figure 4.2: Predictions of March 2019 for Company B

4.6 Combined Method

We believe that after finishing with different testing approaches considered for this study which will be investigated in the next chapter, we can optimize our prediction results from the accuracy perspective utilizing a mathematical model called weighted average model. In this method different variables get different weights considering the importance and the effect that each of them can have on the ultimate results and then an average value is calculated. Accordingly we will first test the accuracy of the algorithms and then we will choose two of the best with the highest results. Afterwards we will calculate the values of each month of 2019 utilizing Equation 1 with the two selected algorithm results. We will call this method "Combined Method" or CM and thereupon demonstrate the comparison results with other algorithms.

Equation 1:

$$\text{Average Weighted Value} = \frac{(\text{Variable1} \times \text{Related Weight}) + (\text{Variable2} \times \text{Related Weight}) + ...}{\text{Total of Weights}}$$

We will demonstrate and analyse all the prediction results for each month of 2019 related to both companies in the next chapter. Also we will compare the prediction capability of each algorithm and the combined method.

Chapter 5

Analysis and Results

As proposed by Hevner [2] in an information systems research study "The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods" and that the accuracy of an IT-Artifact can be evaluated as the quality attribute. Since design is an iterative and progressive process, feedback from the evaluation phase plays a key role for the final artifact to be satisfactory for the problems it was meant to solve [2].Therefore, after finishing with the development of our prototype we need to test it rigorously to be able to analyze the prediction results. In order to answer RQ2 we investigate our results from two different approaches. First we compare the four machine learning algorithms with the closest prediction criterion. Afterwards we investigate the percentage of accuracy of each month prediction by each of the algorithms.

5.1 First Approach: Number of Closest Predictions in one year

In this section we will investigate two tables including all the available real data from the beginning of the year 2014 to the end of 2019.

5.1.1 Predictions of Company A

In the figure 5.1 we show all the real values of data from company A and the results of each month predictions and highlighted the closest predictions to the real values of 2019 with the blue color. As illustrated in this figure SVR_RBF has the highest number of closest predictions with 8 close predictions out of 12 which becomes 66.6% of close predictions in year 2019.

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Real Data (Mins)	2014	10.2	9.6	9.4	11.1	10.5	12.5	13.2	13	10	9.1	10.3	11
	2015	9.5	9.3	10.3	12.2	13.2	14	14.3	12.7	10.9	9.3	9.6	12.6
	2016	10.3	7.5	11	8.2	11.8	15.1	14.9	12	9.9	9.8	7.8	12
	2017	12.5	10.8	10	11.1	12.2	14	16.1	14.9	15	12	8.9	17.1
	2018	11	12.4	14.1	12.5	15.2	15.8	17.3	15.5	14.9	12.7	10.3	13.2
	2019	12	10.1	11.2	11.4	12.3	16.1	17.2	14.3	13.1	11	7.6	11
Predictions of 2019	SVR_RBF	13.83	10.80	10.95	11.85	11.57	12.41	15.31	14.55	12.43	12.62	10.7	8.86
	SVR_Lin	13.35	13.23	13.25	13.16	13.19	13.16	13.29	13.34	13.55	13.44	13.47	13.43
	SVR_Poly	14.73	14.19	13.59	13.63	13.53	13.55	13.72	14.06	14.19	14.24	14.09	13.95
	Linear Regresion	13.84	13.78	13.60	13.50	13.42	13.40	13.61	13.87	13.95	13.95	13.84	13.54

Closest Prediction

Figure 5.1: Closest Predictions to 2019 for Company A

5.1.2 Predictions of Company B

In the figure 5.2 we show all the real values of data from company B and the results of each month predictions and highlighted the closest predictions

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Real Data (Mins)	2014	12	10.5	10.7	9.8	12	14.1	14.6	12.8	11.2	8.8	9.2	11.2
	2015	11.1	10	9.2	11	10.1	10.2	12.4	12.6	11	8.2	9.1	10.2
	2016	10.8	8.9	10.1	9.2	11.8	15	14.1	11.3	10	11	8.3	13.1
	2017	11.9	10.2	9.1	10.3	11.5	13.3	17.2	13.6	14.2	12.2	8.1	15.3
	2018	9.8	10.5	12.6	12.5	13.2	16.1	16.9	14.8	13.6	11.6	9.7	12.9
	2019	11.2	12.1	12.9	11.2	15.1	17.8	19	16.3	14.2	11	10.1	13
Predictions of 2019	SVR_RBF	13.65	9.95	13.05	12.05	11.08	15.30	15.03	16.13	12.98	13.01	10.37	11.33
	SVR_Lin	12.76	12.8	12.75	12.83	12.83	12.91	12.99	13.14	13.26	13.36	13.35	13.31
	SVR_Poly	13.33	13.33	13.25	13.08	13.15	13.26	13.58	13.88	14.14	14.24	14.31	14.28
	Linear Regresion	12.83	12.76	12.76	12.8	12.74	12.92	13.25	13.64	13.85	13.92	13.81	13.65

Closest Prediction

Figure 5.2: Closest Predictions to 2019 for Company B

to the real values of 2019 with the blue color. As illustrated in this figure SVR_RBF has the highest number of closest predictions with 6 close predictions out of 12 which becomes 50% of close predictions in year 2019.

5.1.3 Comparison of the Closest Predictions

As it is shown in figure 5.3 the comparison bar chart shows the percentage of number of closest predictions for all the 12 months of 2019. We can see that SVR_RBF algorithm has the highest rate of closest predictions for both companies A and B. As illustrated bellow in the number of closest predictions by percentage with SVR_RBF for the companies A and B respectively are 66.6% and 50% and with SVR_Lin respectively are 16.7% and 25% and with SVR_Poly respectively are 16.7% and 16.7% and for Linear Regression

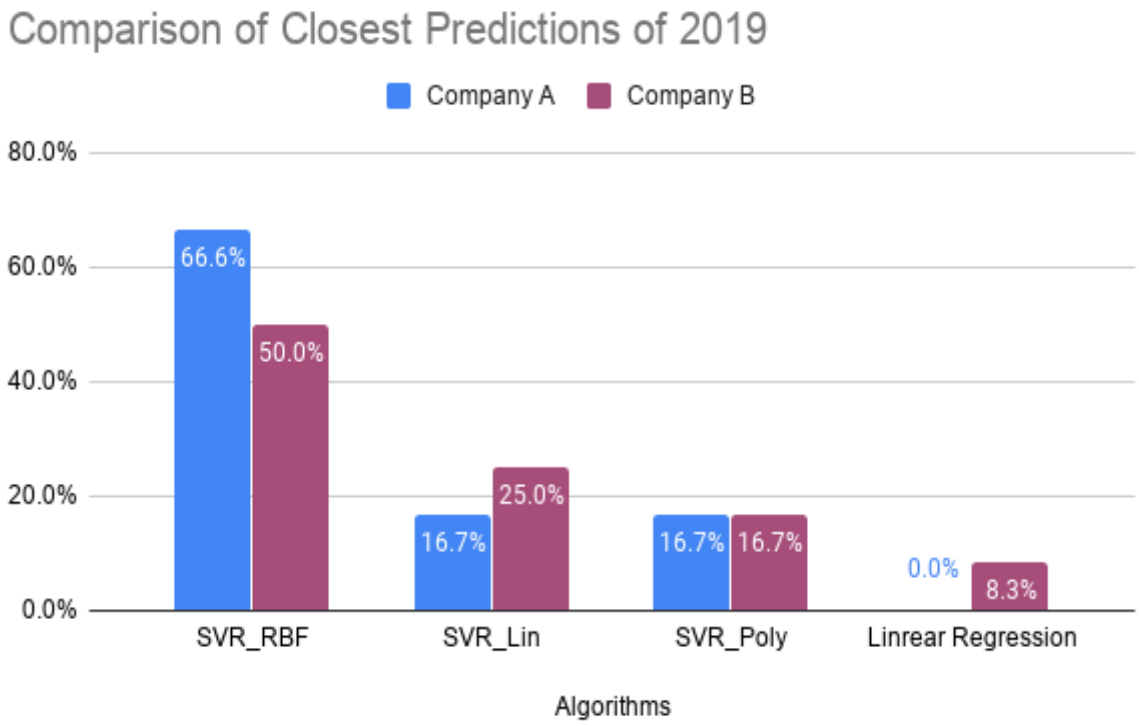

Figure 5.3: Comparison of the Closest Predictions

respectively are 0% and 8.3% which makes the SVR_RBF predictions the closest.

5.2 Second Approach: Prediction Accuracy

In order to calculate the accuracy of each prediction we have used Equation 2 [20] In this section we will have two tables with which we have calculated percentage of prediction error and percentage of prediction accuracy for all the prediction values previously shown in figures 5.1 and 5.2 In the following tables we call prediction error as PE and prediction accuracy as PA and average of each row as Avg.
Equation 2 [20]:

$$\text{Percentage Prediction Error} = \frac{\text{measured value} - \text{predicted value}}{\text{measured value}} \times 100 \text{ or}$$

$$\text{Percentage Prediction Error} = \frac{\text{predicted value} - \text{measured value}}{\text{measured value}} \times 100$$

5.2.1 Prediction Accuracy table for Company A

Below is the figure 5.4 including all the prediction errors and prediction accuracy and their average (Avg column) for each algorithm being evaluated related to company A.

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
SVR_RBF	PE	15%	6.9%	2.2%	3.9%	5.9%	22.9%	10.9%	1.7%	5.1%	14.7%	40.7%	19.4%	12.4%
	PA	85%	93.1%	97.8%	96.1%	94.1%	77.1%	89.1%	98.3%	94.9%	85.3%	59.3%	80.6%	87.6%
SVR_Lin	PE	11.2%	30.9%	18.3%	15.4%	7.2%	18.2%	22.7%	6.7%	3.4%	22.1%	77.2%	17.8	20.9%
	PA	88.8%	69.1%	81.7%	84.6%	92.8%	81.8%	77.3%	93.3%	96.6%	77.9%	22.8%	82.2%	79.1%
SVR_Poly	PE	22.8%	40.4%	21.3%	19.5%	10%	15.8%	20.2%	1.6%	8.3%	29.4%	85%	22.3%	24.7%
	PA	77.2%	59.6%	78.7%	80.5%	90%	84.2%	79.8%	98.4%	91.7%	70.6%	15%	77.7%	75.3%
Linear Regresion	PE	15.3%	36.4%	21.4%	18.4%	9.1%	16.7%	20.8%	3%	6.4%	26.8%	82.1%	18.7%	22.9%
	PA	84.7%	63.6%	78.6%	81.6%	90.9%	83.3%	79.2%	97%	93.6%	73.2%	17.9%	81.3%	77.1%

Figure 5.4: Prediction accuracy percentage - Company A

As shown above the average of PE and PA with algorithm SVR_RBF respectively are 12.4% and 87.6%, with SVR_Lin respectively are 20.9% and 79.1%, with SVR_Poly respectively are 24.7% and 75.3% and with Linear Regression respectively are 22.9% and 77.1%.

5.2.2 Prediction Accuracy table for Company B

Below is the figure 5.5 including all the prediction errors and prediction accuracy and their average (Avg column) for each algorithm being evaluated related to company B.

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
SVR_RBF	PE	21.9%	17.7%	1.1%	7.5%	26.6%	14%	20.8%	1%	8.5%	18.2%	2.6%	12.8%	12.7%
	PA	78.1%	82.3%	98.9%	92.5%	73.4%	86%	79.2%	99%	91.5%	81.8%	97.4%	87.2%	87.3%
SVR_Lin	PE	13.9%	5.7%	1.1%	14.5%	15%	27.4%	22.7%	19.3%	6.6%	21.4%	32.1%	2.3	15.1%
	PA	86.1%	94.3%	98.9%	85.5%	85%	72.6%	77.3%	80.7%	93.4%	78.6%	67.9%	97.7%	84.9%
SVR_Poly	PE	19%	10.1%	2.7%	16.7%	12.9%	25.5%	20.2%	14.8%	0.4%	29.4%	41.6%	9.8%	16.9%
	PA	81%	89.9%	97.3%	83.3%	87.1%	74.5%	79.8%	85.2%	99.6%	70.6%	58.7%	90.2%	83.1%
Linear Regresion	PE	14.5%	5.4%	1.1%	14.2%	15.6%	27.4%	20.8%	16.3%	2.4%	26.5%	36.7%	5%	15.4%
	PA	85.5%	94.6%	98.9%	85.8%	84.4%	72.6%	79.2%	83.7%	97.6%	73.5%	63.3%	95%	84.6%

Figure 5.5: Prediction accuracy percentage - Company B

As shown above the average of PE and PA with algorithm SVR_RBF respectively are 12.7% and 87.3%, with SVR_Lin respectively are 15.1% and 84.9%, with SVR_Poly respectively are 16.9% and 83.1% and with Linear Regression respectively are 15.4% and 84.6%.

5.2.3 Comparison of the Prediction Accuracy

In this section as shown bellow in figure 5.6 we can compare by precise numbers that the SVR_RBF is the most precise algorithm for predicting flight delays for both companies compared to other algorithms utilized in our artifact and considering the amount of data that was available to us for

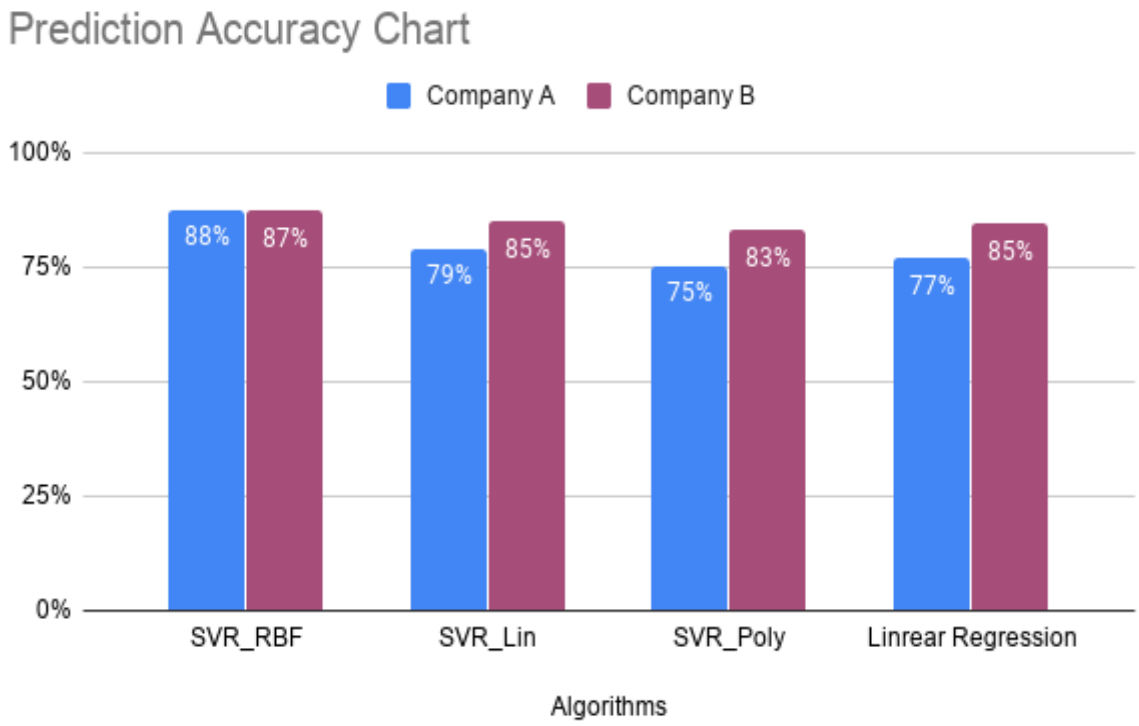

Figure 5.6: Prediction Accuracy of the Algorithms

the testing purposes. We have used the Avg column of figures 5.4 and 5.5 to create figure 5.6 prediction accuracy chart.

As illustrated above in Figure 5.6 the prediction accuracy by percentage with SVR_RBF for the companies A and B respectively are 88% and 87% and with SVR_Lin respectively are 79% and 85% and with SVR_Poly respectively are 75% and 83% and for Linear Regression respectively are 77% and 85% which makes the SVR_RBF the most precise prediction algorithm amongst all.

5.3 Combined Method

We have created a method by combining the results from two of the best prediction algorithms in our artifact which according to the figures 5.6 and

5.3 are SVR_RBF and SVR_Lin and by utilizing weighted average method as explained in the previous chapter equation 1 between the two algorithms' results we will have new results for each month. We have tested many numbers to investigate and discover the most proper weights for each of the selected algorithms to create Equation 3 for the purpose of optimizing our results. Our proposed method to calculate values for each month is Equation 3 as follows:
Equation 3:

$$\text{Average Weighted Value for each month} = \frac{(\text{SVR\_RBF Result} \times 0.9) + (\text{SVR\_Lin Result} \times 0.7)}{1.6}$$

Then after we calculate each month value with equation above, we calculate the prediction error (PE) and prediction accuracy (PA) for each of the values compared to the real values of 2019 for both companies using equation 2. Afterwards we calculate the average percentage of accuracy for this new method and the new results are shown in figure 5.7 in which the CM column is calculated with equation 4 as follows:
Equation 4:

$$\text{CM} = \text{Average(PA of weighted values)}$$

5.3.1 Results for Weighted Average Method

In the figure 5.7 we show the results of each month weighted average value and percentage of accuracy (PA) and the average of PA as results of the CM method for all months of 2019 related to both companies. Thus, as illustrated below in figure 5.7 the CM method's prediction accuracy results for companies A and B respectively are 92.61% and 86.50%.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	CM
Company A Real Values	12	10.1	11.2	11.4	12.3	16.1	17.2	14.3	13.1	11	7.6	11	
Average Weighted values	13.62	11.86	11.95	12.42	12.27	12.73	14.42	14.02	12.92	12.97	11.91	10.85	
PA	86.5	82.54	93.24	91.02	99.82	79.11	83.87	98.04	98.62	82.01	43.26	98.72	92.61%
Company B Real Values	11.2	12.1	12.9	11.2	15.1	17.8	19	16.3	14.2	11	10.1	13	
Average Weighted values	13.26	11.19	12.91	12.39	11.84	14.25	14.13	14.82	13.1	13.16	11.67	12.19	
PA	81.6	92.53	99.85	89.36	78.44	80.08	74.40	90.93	92.27	80.33	84.41	93.81	86.50%

Figure 5.7: Combined Method (CM) Results

5.3.2 Comparison of prediction Accuracy for the CM

As it is shown in figure 5.8 the CM method utilized for company A has been optimized by 5% as we rounded the percentage numbers for all the accuracy comparison charts in this study. However, this method do not demonstrate any significant change to the prediction accuracy for company B.
As illustrated bellow in Figure 5.8 the prediction accuracy by percentage with CM method for the companies A and B respectively are 93% and 87%, with SVR_RBF for the companies A and B respectively are 88% and 87%, with SVR_Lin respectively are 79% and 85%, with SVR_Poly respectively are 75% and 83% and with Linear Regression respectively are 77% and 85% which makes the CM method an optimized method for the prediction accuracy purpose.

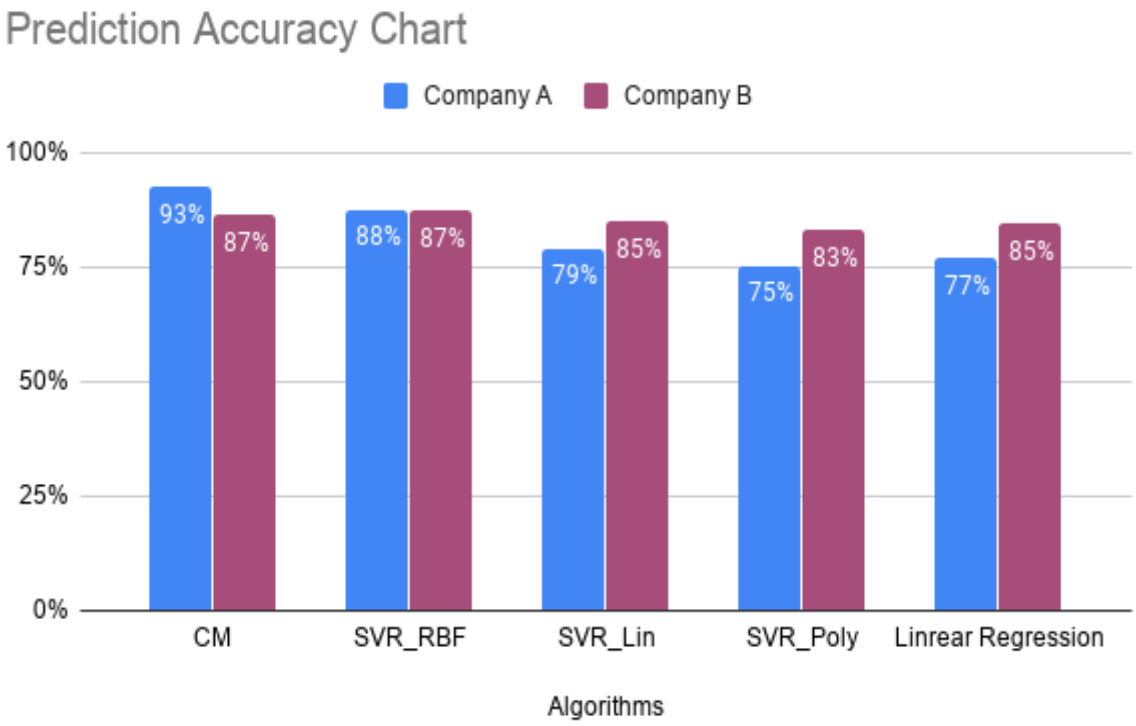

Figure 5.8: Comparison of All methods

5.4 Threats to validity

- Results of this project could be disproved in the times of international crisis or pandemic situations like the one the world is experiencing at the time we are writing this report with corona virus because of the pandemics' direct effect on the airline industry.
- Considering the limited amount of data that we had access to including 6 years of flight delays data related to two airline companies, the results analysis and optimizations are being done. In order to have more accurate results, utilizing more test cases and more datasets will help further investigations to be more efficient.

5.5 Challenges

The most important challenge we encountered developing this artifact was the limitations in order to access the data we needed to test it. According to the regulations between Aviolinx and their customers which the data is related to, we ultimately were authorized to use 6 years flight delay data from two airline companies from 2014 to 2019 anonymously and we did accordingly.

5.6 Contribution

This research study demonstrates evaluation and investigation of the prediction capabilities of the presented algorithms and proposed method for the flight delays with the specific focus on slot scheduling efficiency. From the business perspective, the contribution of this study is that it can be utilized for adding machine learning prediction capability with the investigated methods to the slot scheduling system of the aviation software products in order to reduce delays with an insight into future delays. Also from the research community perspective, analysis, the comparison results and the methods presented can be utilized for more investigations in order to develop more accurate prediction methods for flight delays to use into slot scheduling systems.

Chapter 6

Conclusion and Future Work

As explained earlier in this report, since airport slots are considered as an asset to airline companies, losing them in an airport directly affects them financially. So companies should take any necessary step towards more efficient usage of slots not to lose them according to 80-20 Grandfather rights (GFR). This project can contribute to their software systems by utilizing machine learning techniques in order to predict future flight delays. We have answered our RQ1 in chapter 4 of this research study by illustrating how raw flight delays data can be converted into usable future delays predictive information with which airline softwares could add more efficiency to their scheduling systems. Afterwards we have answered our RQ2 by analyzing the results and demonstrating the comparison between the previously presented prediction methods in chapter 5 of this research study. As the comparison charts demonstrated the accuracy of utilized algorithms in the previous chapter, the most powerful and reliable prediction algorithm amongst the ones investigated in the prototype we created is SVR_RBF with highest rates of both accuracy and closest predictions. Afterwards we tried to propose a new model utilizing mathematical weighted average model in order to optimize the prediction accuracy results which we called it combined model or CM and as the results analysis demonstrated the CM method showed higher overall accuracy for company A and the same results for company B as we rounded the numbers in the last comparison chart in figure 5.8. Therefore, by utilizing

the prediction results from this research paper and the methods proposed, within slot messages, airport coordinators and airline companies would be able to plan their schedules in a more competent state and efficacious way. Future investigations are necessary with larger amounts of data, more test cases and more assessments of the methods proposed in this research study. The correlation between the data from larger datasets may result in different weights in our CM method. All in all, the more development and optimization accomplished in this area, the more economically lucrative airline information systems become.

References

[1] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[2] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[3] Mamello Thinyane. Small data and sustainable development — individuals at the center of data-driven societies. 2017. Publisher:https://ieeexplore.ieee.org/document/8246991.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Konstantinos Zografos, Michael Madas, and Konstantinos Androutsopoulos. Increasing airport capacity utilisation through optimum slot scheduling: Review of current developments and identification of future needs. *Journal of Scheduling*, 09 2016.

[6] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong, and F. Xia. A survey of scholarly data visualization. *IEEE Access*, 6:19205–19221, 2018.

[7] Chun-YangZhang C.L.Philip Chen. Data-intensive applications, challenges, techniques and technologies: A survey on big data. 2014. Publisher:ScienceDirect , https://www.sciencedirect.com/science/article/pii/S0020025514000346?via

[8] P. P. Shinde and S. Shah. A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6, 2018.

[9] S. Angra and S. Ahuja. Machine learning and its applications: A review. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pages 57–60, 2017.

[10] P. Louridas and C. Ebert. Machine learning. *IEEE Software*, 33(5):110–115, 2016.

[11] S. Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39, 2019.

[12] J. Wang, X. Chen, and S. Guo. Bus travel time prediction model with support vector regression. In *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pages 1–6, 2009.

[13] Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio Trani, Bo Zou, Rodrigo Britto, Doug Fearing, Prem Swaroop, Nitish Uman, Vikrant Vaze, and Augusto Voltes. Total delay impact study: A comprehensive assessment of the costs and impacts of flight delay in the united states. 10 2010.

[14] John V. Guttag. *Introduction to Computation and Programming Using Python: With Application to Understanding Data (The MIT Press)*. The MIT Press, 2016.

[15] T. Carneiro, R. V. Medeiros Da NóBrega, T. Nepomuceno, G. Bian, V. H. C. De Albuquerque, and P. P. R. Filho. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, 6:61677–61685, 2018.

[16] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[17] Travis Oliphant. NumPy: A guide to NumPy. USA: Trelgol Publishing, 2006–. [Online; accessed ¡today¿].

[18] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[19] Xin Yan and Xiao Gang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., USA, 2009.

[20] Wu Guang, Massimo Baraldo, and Mario Furlanut. Calculating percentage prediction error: A user's note. *Pharmacological Research*, 32(4):241 – 248, 1995.