Aalto-yliopisto
Aalto-universitetet
Aalto University

Bachelor's Programme in Data Science

COVID-19's impact to Finland taxi industry

Dylan Nguyen

Bachelor's Thesis
2020

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Dylan Nguyen

Title of thesis COVID-19's impact to Finland Taxi Industry

Programme Bachelor Program of Data Science

Major Data Science

Thesis supervisor William Wilkinson

Date 21.12.2020 **Number of pages** 27 + 3 **Language** English

Abstract

This study traces the impact of COVID-19 pandemics to the taxi industry in Finland. A 14-year subset of taxi shifts data is provided thanks to Trafore Oy. During this study, COVID-19 pandemics was estimated using indicators such as Google Trend search interest for keyword "COVID", as well as the daily new cases of COVID-19 in Finland. Regarding the taxi industry, taxi revenue per shifts and number of shifts are used as a performance indicator. The study found a strongly negative correlation between taxi revenue per shift and Google search interest. On the demand side, the taxi revenue per shift experienced a staggering drop of 70% within March but has gradually increased since. However, taxi revenue per shift but failed to reach the rate of previous years. The impact of COVID-19 is the most apparent from March to June, and from September onwards, with March and April being the most affected months in terms of taxi revenue per shifts. With the obtained insights from the study, an attempt to predict the taxi revenue rate was made using Prophet. However, the upcoming future of the taxi industry is difficult to predict, with disruptive changes on the horizon, such as the end of COVID-19 pandemics and further changes to the taxi regulation.

Keywords Data Exploration, Data Mining, Big Data, Taxi Industry, COVID-19

Preface

This thesis would not be possible without the generosity and insightful comments of Ville Ranta, who is my direct supervisor at Trafore Oy and who provided the dataset for this paper. I would like to thank my instructor William Wilkinson for his guidance and patience through the writing process. I'm deeply indebted to my long-time friend and co-worker Anh Ta, whose valuable advices and extensive knowledge cannot be overestimated. And last but not least, I would like to thank my girlfriend Van Anh, who has never wavered in her support and trust in me.

Espoo, 21.12.2020

Dylan Nguyen

Table of Contents

1 Introduction

The arrival of COVID-19 virus in spring 2020 has been causing a significant reduction in social events and activities, which immediately affected global economies and posed great challenges for a lot of businesses and industries, including taxi industry in Finland. According to Alusta [1], taxi companies approximate 80-90 % decline in amount of taxi rides within a week between 13-20 March. Finland's Taxi Association demanded responsibility in terms of income compensation from municipalities and the state for taxi entrepreneurs, who suffer from the suddenly ended school transport as one major aspect of the reduced mobility [16].

The purpose of the thesis is to utilize data mining and exploration tools and techniques to provide Trafore Oy insights into their multidimensional taxi shift data and explore the impact of COVID-19 pandemic have on the Finland taxi industry, as well as on the number of Trafore Oy customers as taxi operators, which are presented in the provided dataset.

Various time series correlation statistics are utilized to estimate the correlation as well as the causation relationship between COVID-19 cases in Finland and taxi revenue data. Afterwards, the collected insights are utilized to predict the taxi shift revenue in the near future, using Facebook Prophet as a prediction model.

The paper seeks to provide answers to the following questions:

- Has COVID-19 pandemics had any impact to the Finnish taxi industry?
- If yes, how big is that impact?
- What is the future of Finnish taxi industry after COVID-19?

In what follows, the second chapter thoroughly reviews the data exploration techniques and related studies. Chapter 2 presents the relevant literature about taxi industry in Finland, as well as various methodologies used in the paper. Chapter 3 presents the overview of the provided taxi shifts data, from Trafore Oy, and COVID-19 dataset, as well as necessary pre-processing. Chapter 4 and 5 present and analyse the findings. The last chapter offers concluding remarks as well as discusses possible limitations and further studies.

2 Literature review

This section introduces relevant literature about the research topics, which are the de-regulation of Finnish taxi service, and recent studies about COVID-19 epidemic on the global traffic industry, as well as Finnish taxi industry. The section also explores methodologies which were utilized in the paper, which include time series correlation coefficients such as Pearson Correlation, and Cross Correlation. Furthermore, Facebook Prophet was the main prediction model, due to its simplicity as well as flexibility to incorporate extraneous regressors into predicting the main dependent variable.

2.1 Taxi Industry in Finland

This section discusses the taxi de-regulation in July 2018, whose impact to Finnish taxi industry is compared with COVID-19 pandemics [6]. Before July 2018, the industry had been highly regulated by the government in terms of prices and availability. Some of the noteworthy regulations [13] are as followed:

- Taxi fare was similar across the country, regardless of locations.
- Each municipality was granted a maximum number of taxis, thus preventing taxi companies to simply provide service on urban areas and disregard taxi service in rural areas.
- For eligibility to be a taxi driver, one needs to take a compulsory course, on top of a driving examination and a health test.
- Once registered to a geographical area, taxis are restricted and can only provide taxi services to that specific area.

After July 2018, disruptive changes were introduced, which made the industry open for competition. Taxis are no longer restricted to a geographical area and are allowed to provide taxi service anywhere in Finland. Pricing policies can be set flexibly by taxi entrepreneurs instead of government's previously regulated prices [13]. These changes have re-introduced ride-hailing companies back to the market. Since the dispute against the government in regard to unlicensed taxi drives [31], Uber has suspended their services in Finland and wait for the deregulation to take effect. Thanks to removal of taxi licenses limitation and fare restriction, Uber officially relaunched on 03-07-2018 [30], and has been doing relatively well in terms of market share. Statista [32] shows that among the ride-hailing application in 2019, Uber accounted for 20% of the market. Yango, another ride-hailing application which was launched later in November 2019, accounted for 10%. On the contrary, traditional taxis

have also adopted to the wave of ride-sourcing technology and brought forth their own applications to compete with foreign companies, such as Lähitaksi and Taksi Helsinki, whose percentage in the Finland ride-hailing market is 20% and 15%, respectively.

2.2 Exim.fi as a service

The thesis' idea is based on and inspired by Exim.fi, which is a taxi information management service, supporting both website and desktop application. Exim.fi is owned by Trafore Oy, who provided the dataset for this study.

The data pipeline of Exim.fi is illustrated in figure 1. Firstly, customers order taxis either by phone, or by online, such as on the website and mobile application booking systems. Afterwards, trips are assigned to taxi drivers. When a trip is completed, the taxi meter in the car sends the information about the trips to taxi meter services, such as start and end timestamps, payments, payment methods, etc. Exim.fi frequently synchronizes the data from these servers to store and manage the data with the in-house server solution. From the taxi companies' point of view, Exim sorts the data based on the end user's results and present the data in an easy-to-comprehend and easy-to-understand format, such as tables and pdf reports. Users can create monthly salaries for their drivers, invoices for their customers, synchronize data into different endpoints, for example salary reporting services and Incomes Register.

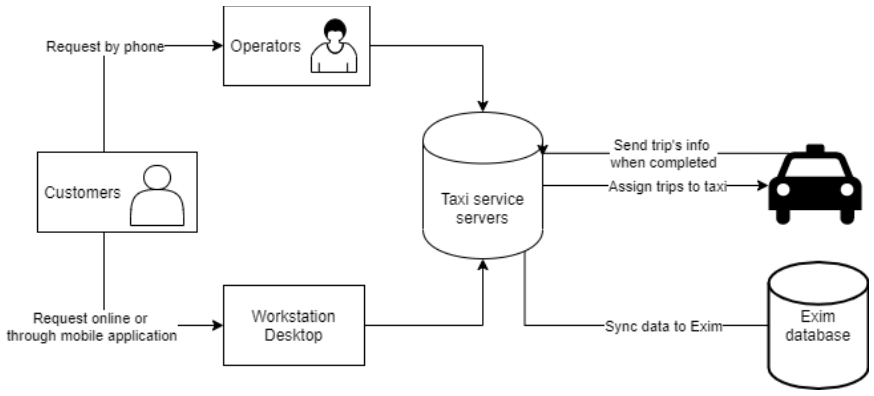

Figure 1: Data pipeline of Exim.fi, from customer to database

2.3 Global impact of COVID-19 on the taxi industry

Since the arrival of the pandemics, there has been a staggering amount of attention from the science community, investigating the impact of COVID-19 to the traffic and transportation industries, both locally and globally. Tahlyan and Mahmassani [5] explored traffic and

collision data in Chicago during the COVID-19 crisis and found that the number of crashes decreases, but the severity of resulting injuries actually increases. Mahmassani co-authored another paper with Ale-Ahmad [9] exploring the same dataset and showed the lockdown order in Chicago started in March reduced taxi ridership by 95% and the number of operating taxis by 85%. Hu et al. [20] utilized a Bureau of Public Road function [14] to correlate travel time and number of commuting cars. They found that travel time increases of 5-10 minutes are possible in high-transit cities, which adds up to hundreds of thousands of hours of additional travel time each day. These increases are avoidable if transit ridership resumes in step with car traffic. After conducting a spatiotemporal analysis of taxi demand using trajectory data in Shenzhen, China, Hongyu et al [33] found that the taxi demand in Shenzhen shrank by 85% in the lockdown phases and barely recovered when the city reopened. Furthermore, the authors found that the taxi companies adapted to these changes by cutting back working hours and adjusted the schedule to serve peak hour periods.

2.4 Facebook Prophet

Prophet [34] is a time series prediction model from Facebook, which is henceforth referred as Fbprophet. Its advent has solved two main issues in extrapolating a time series in practice: the inflexibility of automatic models in parameter tuning; and the lack of professional skills for the analysts to produce high-quality time series prediction.

Facebook Prophet implements a Generalized Additive Model, and - in a nutshell - models a time-series as the sum of different components (non-linear trend, periodic components and holidays or special events) and allows to incorporate extra-regressors (categorical or continuous).

Theoretically, Prophet represents the data as being made up of several components, following the equation:

$$\mathbf{y(t) = g(t) + s(t) + h(t) + \varepsilon_t}$$

g(t): The trend function that models non-periodic changes exploiting two trend models. By default, the piecewise linear regression model with changepoints is used for the forecasts. When the problems exhibit saturating growth, we can switch to a nonlinear model with logistic growth and set maximum and minimum achievable points.

s(t): The periodic changes (yearly, monthly, daily seasonality) based a partial Fourier sum. We can disable certain kinds of seasonality and change the frequency changes.

h(t): The effect of holidays and special events. We can add manually a set of holidays and events or use built-in data from the library. The prior scale could be reduced to dampen the holiday effect.

Besides, Prophet's measurements do not require pre-processing missing values or outliers. There was also wide space for tunings, such as additional regressors or adjusting prior scales, which significantly improved our forecasting accuracy. In this study, Prophet prediction model is fine-tuned by experimenting with different set of additional regressors as well as hyperparameters.

2.5 Correlation and synchrony metrics

During this sub-section, the methods of correlation and synchrony between individual time series are explored, starting from naïve methods, such as Pearson correlation, to advanced methods that explore more fine-grained dynamic interaction between two signals, such as Cross Correlation.

Pearson correlation measures linear relationships between variables. It assumes that the variables are distributed normally. Pearson's correlation is calculated by dividing the covariance of two variables by the product of their standard deviation. Covariance measures how two variables move together over time. When we divide the covariance by the standard deviation, we make the Pearson correlation unit smaller and therefore always between the values -1 and 1. The formula for Pearson Correlation is as followed:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Figure 2: Mathematical formula for Pearson Correlation

The biggest limitation of Pearson's correlation is that it assumes that the variables have a linear relationship between them. Most variables do not have a linear relationship. Financial assets, for example, have a non-linear relationship between them.

When the value of Pearson's correlation is zero, it means there is no linear relationship between the two variables. However, there may be a non-linear relationship between the variables. Hence a value of 0 does not mean that the two variables are completely independent of each other.

Pearson Correlation is a good place to start, however will not provide a whole picture of time series dynamics and how one has a lagging effect on another, which can be measured by

cross correlation. Cross correlation is a more advanced version of Pearson Correlation. Instead of using Pearson Correlation on two time series and looking naively at the correlation of the two, cross correlation incrementally shifts on time series at each step and calculate correlation coefficients of the shifted version of the first time series and the second time series. This provides a more in-depth analysis into the dynamics of two time series: a leader-follower relationship, where the leader initiates an event, which the follower repeats. As the correlation is looked from different lags, this dynamic is explored more thoroughly.

In order to identify whether the correlation between two time series is statistically significant, permutation test is utilized. The steps in performing a permutation test are listed as followed:

- The time series observations are coupled randomly.
- Pearson correlation coefficient is calculated for this permuted sample.
- Repeat this process for a significant number of times, for example 999 times.
- Order the obtained 1000 (the original plus the ones calculated from permuted obs) Pearson correlations and check if the original is on the tail (for example does it belong to the largest 5%). If that holds, the null hypothesis (no linear dependence) is rejected and state that there is linear dependence.
- Adjust the p-value if different tests on different lags are performed

There may be correlation with lag 0, but no correlation with lag 1 or vice versa or both could be significant, or both could be insignificant. This approach does not require normality assumptions from both time series, as compared to naïve Pearson correlation.

3 Exploratory Data Analysis

This chapter describes the data in question, how they were pre-processed and investigates the dataset using multiple exploratory data analysis (EDA) techniques. The dataset was provided by Trafore Oy, which is a subset of the actual taxi shifts database at Exim.fi. The dataset contains millions of rows. Each row represents a taxi working shift, which are collected from 2007 until the November 2020. The columns in the dataset include information about taxi company, driver, car identification, start time and end time of a shift, revenue in Euros before and after VAT. Each shift contains multiple taxi trips, which a driver picks up a single passenger or a group of passengers and transport the customer(s) to their desired location. Due to a confidentiality agreement with the company, axis of some graphs and figures are removed.

In this study, the main variable in question is the aggregated revenue in Euros per taxi shift including VAT (henceforth to be referred as TR), and the number of taxi shifts (henceforth to be referred as NS). TR is calculated by dividing the sum of revenue by the number of shifts. Both the TR and NS are aggregated either daily, monthly or yearly. In order to clean up the data, several methods have been applied to the dataset, which is listed below:

- Removing rows with outlier values in the start date and end date attributes. Several rows have invalid date attributes and was removed from the dataset.
- Removing rows with outlier values in the revenue attributes. Up to 10 shifts have revenue over 10 million Euros, which massively skewed the average. The 99.99 quantile of revenue attribute was obtained, and all the shifts which have revenue higher than that was removed from the dataset.

The number of rows was reduced by approximately 15% after pre-processing and removing outliers. Figure below reports the taxi revenue per shift from 2007 to 2020, aggregated by year and month.

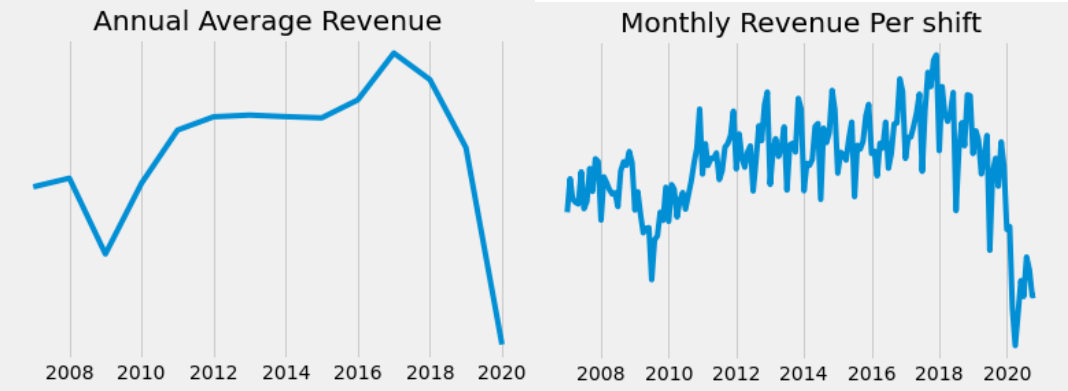

Figure 3: Annual (left) and Monthly (right) average of taxi revenue per shift

Compared to historical averages, we can see a significant drop in revenue in 2020 due to COVID-19. The lowest point in revenue per shift average in March 2020 is observed to be all-time low in the observed data, even lower than the 2009 average revenue, which was possibly due to the global economic crisis in 2007-2008. The impact of COVID-19 is explored in-depth in later section.

3.1 Data Distribution

The figures show the distribution of revenue for all the data, which is from 2007 to the end of 2019 on the left-hand side, and for the data from 2020 on the right-hand side.

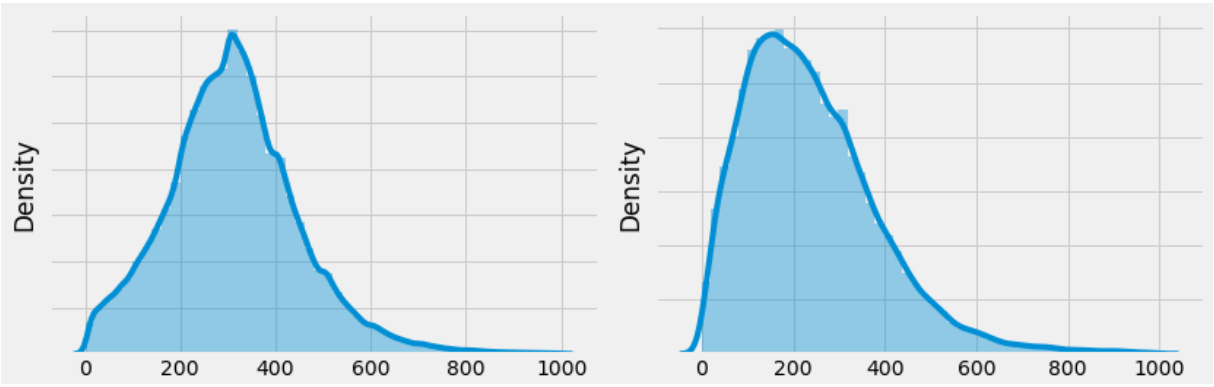

Figure 4: Distribution of taxi revenue per shift before 2020 observations (left) and after 2020 observations (right)

Before 2020, a shift earned approximately 315 Euros, with the first quantile (25%) and the third quantile (75%) are 200 and 400 Euros, respectively. However, during 2020, the distribution experienced a significant shift towards the left, with the arithmetic mean underwent a 25% reduction, dropping to approximately 240 Euros per shift. The first quantile and the third quantile dropped to also 130 and 320 Euros. This reduction is most likely due to the effect of COVID-19 pandemics, which is explored more comprehensively in the later section.

3.2 Seasonality

We explored the inherent dynamics of seasonality, which includes annual and weekly seasonality of taxi revenue in this sub-section.

Regarding weekly seasonality, the figure shows the distribution of revenue aggregated by day of week. As can be confirmed from figure 5, taxi ridership is low at the beginning of the weak, reaches the peak on Saturday, and reduces on Sunday. Average taxi revenue stays approximately at the same level during weekdays. This is in line with the findings from Kamga et al [3], in which the authors explored New York taxi ridership dataset and found similar seasonality patterns for taxi ridership (total number of passengers).

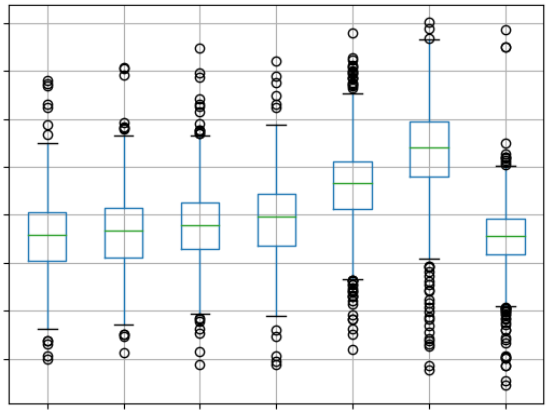

Figure 5: Boxplot of taxi revenue per shift aggregated by day of week, starting on Monday and ending on Sunday

Secondly, the annual seasonality was obtained by removing the weekly seasonality from the original revenue time series to obtain a weekly seasonal-adjusted time series. Afterwards, the time series was decomposed to obtain the annual seasonality component. The purpose of this is to remove the effect of weekly pattern before taking into account of the yearly pattern. Annual seasonality can be seen on the left hand-side and the boxplot for revenue grouped by month on the right hand-side.

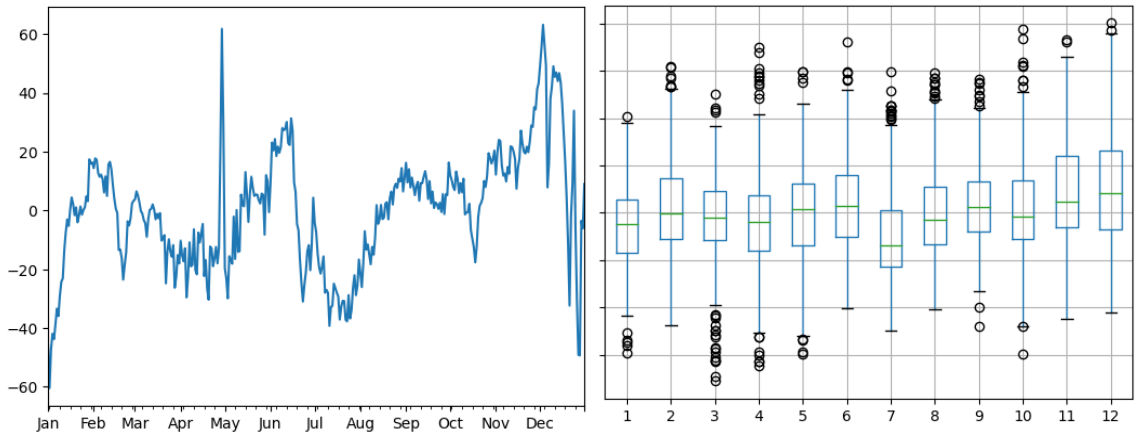

Figure 6: Annual Seasonality for taxi revenue per shift (left) and boxplot of taxi revenue per shift grouped by month (Right)

Some interesting patterns emerge from the first inspection: July and August are slightly less busy months for taxis, when Finnish people tend to leave their homes in the cities to spend their holidays in summer cabins. On the other hand, December is the highest earning month, with the mean of revenue per shift moderately higher compared to the other months in the year. This could be the result of the increase in social events and celebrations related to Christmas and end of year, which leads to increase in usage of taxis, as people usually opt to use taxis instead of driving their own cars.

Several outliers can be reported from the seasonality of the time series as well. Firstly, there is a sudden spike towards the beginning of May, which can be explained by the Finnish holiday Vappu, which is the night between 30.04 and 01.05, when Finnish people gather in a public park for a picnic to enjoy the Mid-summer celebration.

3.3 Google Trend and COVID-19 daily new cases

In order to estimate the impact of COVID-19 to the taxi industry, the correlation of the taxi revenue is calculated against two other time series: Google search interest of the keyword “COVID” in Finland (from now referred as SI), and the daily new cases of COVID-19 pandemics in Finland (from now referred as DC).

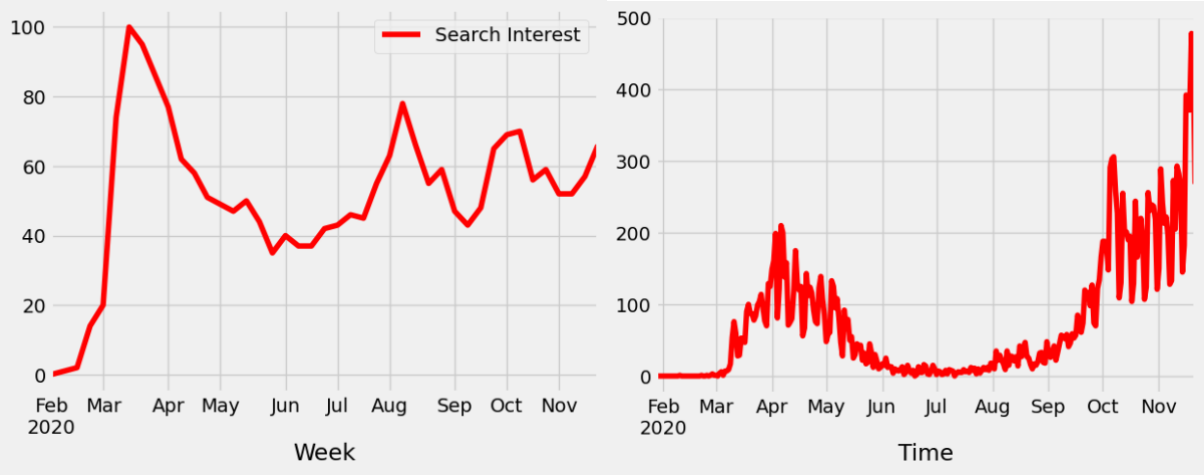

Figure 7: Time series visualisation for search interest dataset (left) and daily new cases of COVID-19 in Finland (right)

The two figures report the temporal distribution of SI (on the left) and DC (on the right). The examine period of SI is from week 4 (20.01. – 26.01.2020) to week 46 (09.11 – 15.11.2020), with the number of observations being $N_1 = 43$ weeks. Data from the Google Trends platform are retrieved in .csv and are normalized over the selected period. Google Trends reports the adjustment procedure as follows: "Search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same search interest for a term don't always have the same total search volumes" [8]. For the specified area and time, numbers reflect search interest relative to the highest point on the graph. The peak popularity for the word is a value of 100. A meaning of 50 indicates that the expression is half as popular. For this word, a score of 0 means there was not enough data. The data may slightly vary based on the time of retrieval.

For the COVID-19 cases data, the examine period for DC is from 28th Jan 2020 to 21st of November 2020, with the number of observations being $N_2 = 299$ days. Data from the distribution of COVID-19 new cases in Finland is retrieved from the official API of Finnish institute for health and welfare [19]. Each observation contains the number of new COVID-19 positive cases in Finland.

4 Correlation with COVID-19 pandemics

This section investigates the correlation of COVID-19 pandemics and the taxi industry in Finland's revenue employing multiple correlation coefficients.

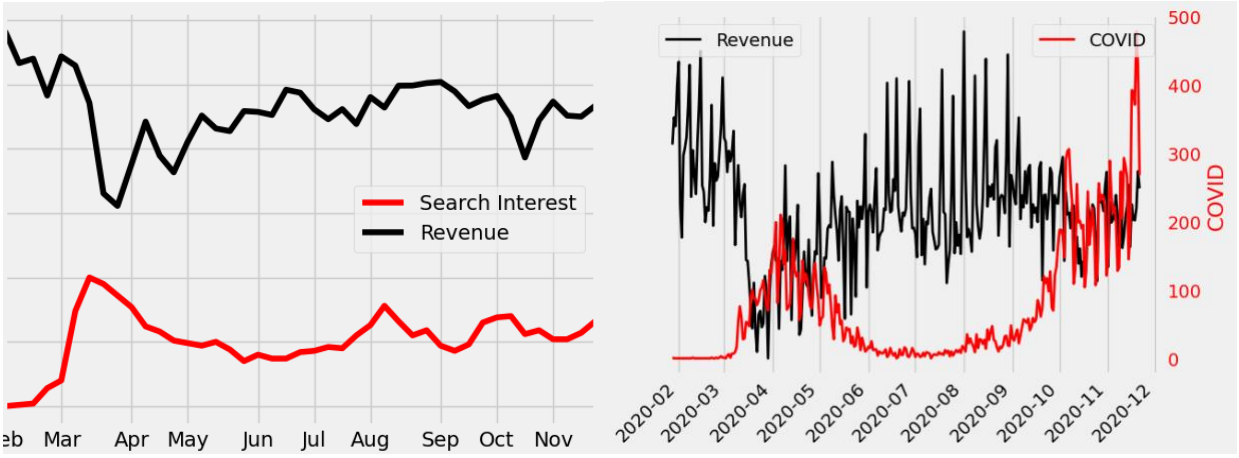

Figure 8: Weekly time series of SI again TR (left) and daily time series of DC against TR (right)

The figure on the left reports the weekly search interest for the keyword "COVID" in Finland (SI), against the weekly taxi revenue (TR), from week 4 to week 46 of 2020. The figure on the right reports daily taxi revenue per shift and daily new cases of COVID-19 (DC) in Finland from January 28, 2020, when the first case of COVID was detected in Finland, until November 21, 2020. TR was aggregated weekly when compared to SI and aggregated daily when compared to DC.

A significant drop of TR is observed on 16.03.2020, which coincides with the announcement of emergency lockdown for the whole Finland from the government, with the number of new cases increased tremendously, as well as SI spiking around March 2020. From week 10 (02.03 – 15.03.2020) to week 11 (16.03. – 22.03 2020), weekly TR dropped by roughly 30%, and the total number of shifts dropped by more than 50% compared to the previous week. During the subsequent weeks, even though TR began to stabilize and gradually increase, the number of shifts continued to reduce. The number of shifts in week 13 (23.03 – 29.03.2020) dropped 70% compared to the number of shifts in week 10, before the first wave of COVID-19 hit Finland. This coincides with the claim from Alusta [1], where taxi companies estimated a significant amount of taxi trips during the week of 13.03 – 20.03.2020.

After the initial plunge in March, TR fluctuated but has steadily increased as shown in both figures. After the total lockdown was lifted on the 01.06.2020, TR remained constant at around the same rate from June. The daily shift revenue figure reported that TR even reached the peak before COVID-19 on several occasions from June to September 2020.

However, the weekly mean of TR was still inferior compared to the weekly mean before COVID. This can be explained by the peak revenue on Saturdays, followed by the reduced revenue for other days of week. However, this is most likely not an impact of COVID-19, since comparing to the rate of previous years, the rate of summer 2020 did not vary by too much. However, from September forwards, the difference between 2020 and the previous years' TR can be observed much clearer, with a clear drop from September 2020 onwards. As can be seen from the percentage change from 2019, March and April are the most affected month, with TR dropped 32%, compared to the same month in 2019. July and August were the least affected, with the percentage change only 11,4 and 17,6, respectively.

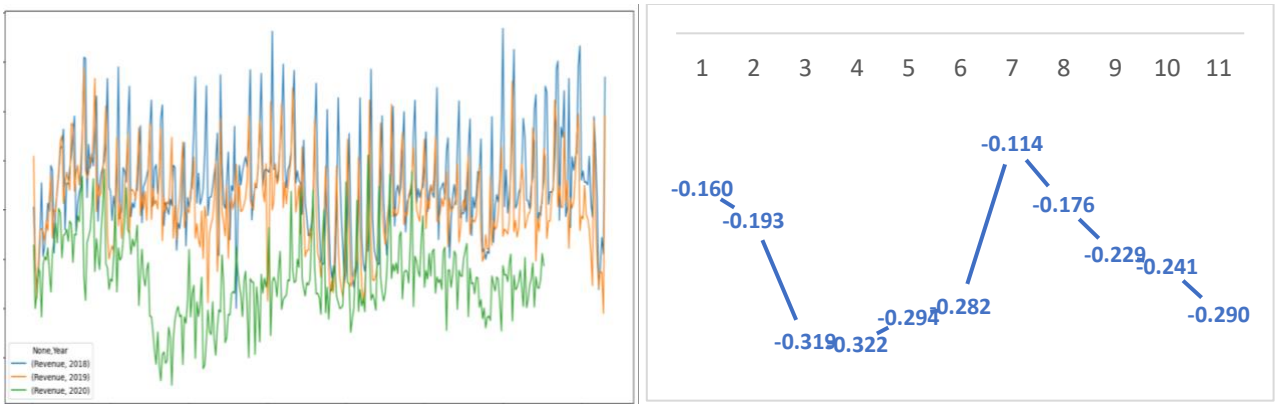

Figure 9: Daily Revenue Per Shift of 2018, 2019 and 2020, (left) and 2020 percentage change from 2019 (right)

During the period from the end of the first wave to the beginning of the second wave, a minor peak is observed in the search interest of COVID, which was in August 2020. This coincides with multiple occasions of COVID-19 positive cases detected, who returned to Finland from abroad after 2020 summer vacation ended [25][27][28][29]. This might have increased worries among the public for upsurge in new infections in addition to a potential second wave.

As the number of new COVID-19 cases peaked and Finland entered the second wave of the pandemic, TR dropped substantially, but on average still remained higher than the March drop. One interesting takeaway is that during the start of both COVID-19 waves, when the number of new cases started to increase, taxi revenue experienced a substantial drop at the beginning of the waves, as can been seen on the middle of March as well as the middle of October, when the number of cases increased considerably at the same time. However, the drops were not sustainable, with TR observed to sharply increase afterwards.

In order to further explore the links between COVID-19 pandemics and the taxi industry, Pearson Correlation Coefficient is calculated. We looked at the correlation between four weekly variables: Search interest for COVID, daily number of cases in Finland, total number of shifts, revenue per shift.

Figure 10: Correlation heatmap for weekly aggregated SI, DC, TR and total number of shifts

From first inspection, the correlation between TR and SI is highly negatively correlated (r=-0.57). Furthermore, permutation test can examine how significant the correlation between TR and SI is. Permutation test is performed by calculating Pearson correlation between a randomly shuffled time series of SI and the original time series of TR, and repeating this calculation 1000 times to form a distribution of coefficients. The original r=-0.57 falls on the tail of this distribution, whose 0.05 tail threshold is -0.25. Thus, the null hypothesis can be rejected, which is there is no linear dependence, and it can be concluded that there is linear dependence between taxi revenue and search interest.

The negative relationship can be reinforced from the scatter plot between SI and TR in figure 11. In general, as the search interest for COVID increases, the revenue decreases, which is also seen from the trendline.

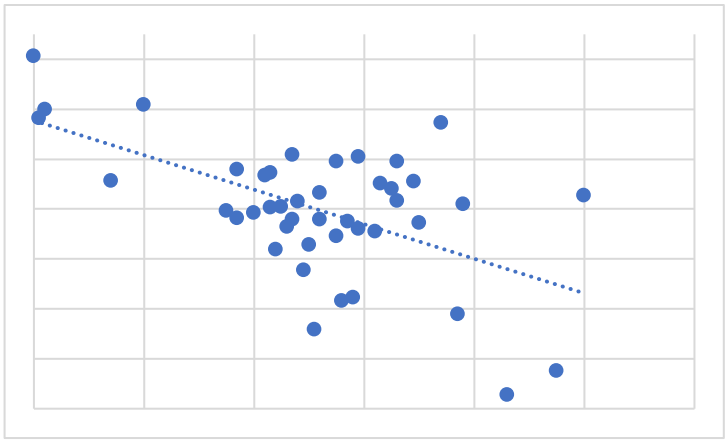

Figure 11: Scatter plot of SI and TR of 2020, with SI on the x-axis and TR on the y-axis

Further exploration into cross correlation between the search interest and taxi revenue also shows the leader-follower relationship between SI and TR, when the correlation between TR and a lagged SI is more highly negative. The highest negative of the cross-correlation function can be found by shifting SI time series by one and two, and the correlation is r=-0.6730 and r=-0.6662, respectively. It can be inferred that SI is leading the interaction, and TR follows with a lag of 1 or 2, which implies that as the search interest for COVID on Google increases, taxi revenue per shift will decrease in one- or two-weeks' time.

5 Prophet prediction

Another approach at isolating the impact of COVID-19 is to fit a predictive model to the data before 2020, and compare the prediction to 2020 revenue, which experienced the impact of COVID-19. The difference between the forecasted value of the model and the actual 2020 revenue can elaborate on the influence of the pandemics to the Finnish taxi industry.

Facebook Prophet was utilized instead of other models, due to its simplicity in implementation, as well as enough complexity to capture seasonality and trend. An attempt to improve the models was made by tuning different hyperparameters and extra regressors. The dataset contains 5050 rows and 6 columns, with each row contains the daily revenue per shift as the dependent variable. The independent variables are date, rain, snow, minimum temperature and maximum temperature. The weather data is retrieved from the Helsinki Kaisaniemi station of Finnish Meteorological Institute, from 01.01.2007 to 21.11.2020. During the training of Fbprophet models, the weather data are added as extra regressors in the Fbprophet model.

The first subsection 5.1 discusses hyperparameters tuning process, and the 5.2 reports on the forecasted value of the best forecasting model and compare the predictions to the actual data.

5.1 Hyperparameter tuning

The examine period of the training set is from 01.01.2007 to 31.12.2018, which contains 4361 observations. The examine period of the testing set is from 01.01.2019 to 31.12.2019, which contains 365 observations. The evaluation benchmarks include mean absolute error (MAE), mean absolute percentage error (MAPE) and Pearson Correlation (r).

Table 1: Formulas of the evaluation metrics

Mean absolute error (MAE)	Mean absolute percentage error (MAPE)	Pearson correlation (r)
$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \lvert y_i - x_i \rvert}{n}$	$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n} \left\lvert \frac{A_t - F_t}{A_t} \right\rvert$,	$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$
y_i = prediction x_i = true value n = number of data points	A_t = actual value F_t = forecast value n = number of data points	x_i =values of x-variable in a sample x_bar = mean of all x values y_i = values of the y-variable in a sample y_bar = mean of all y values

The predictive power of the weather data was measured by training the Fbprophet model on each set of variables in the power set of 4 weather variables (also refereed to as extra regressors). In total, there are 16 distinct sets of the power set of 4 variables, including {{}, {rain}, {rain, snow}, {snow, minimum temperature}, {rain, maximum temperature}, etc}. The purpose is to determine which variables or set of variables add the most forecasting ability to the model. In addition to the extra regressors, the holiday effect is also taken into account. The performance metrics are reported in the table below:

Table 2: Fbprophet model performance comparison with different hyperparameters

Model	*Mean Absolute Error (MAE)*	*Mean Absolute Percentage Error (MAPE)*	*Pearson Correlation*
Vanilla Prophet	37.780	13.281	0.669
Prophet with Holidays	36.711	12.858	0.712
Prophet with weather	37.498 ± 0.399	13.284 ± 0.134	0.672 ± 0.005
Prophet with Holidays and weather	36.571 ± 0.393	12.899 ± 0.131	0.712 ± 0.005
Hyperparameters tuned Prophet Model without manual change point	32.772	11.576	0.724
Final Prophet Model (With Manual Changepoint)	**22.218**	**7.405**	**0.772**

The difference between the models reported in the table are listed as followed:

- For the first model “Vanilla Prophet”, the model was run in its original condition, without any additional holiday information or regressors. The error rate is used as a baseline to evaluate how much predictive power is increased with different hyperparameters.
- For the second row, Finnish holiday is added as an extra regressors. Priors for holiday is set as default.
- For the third row, 16 models were each equipped with a distinct set of weather regressors, but without Finnish holidays. Priors for each hyperparameters are set as default. Mean and standard deviations of error rates of these 16 models are reported.

- For the fourth row, 16 models were each equipped with a distinct set of weather hyperparameters, and with Finnish holidays included. Priors for each hyperparameters are set as default. Mean and standard deviations of error rates of these 16 models are reported.
- The fifth row reports the evaluation metrics of the tuned Fbprophet model, excluding manual change-point.
- The last row reports the final Fbprophet model, including the manual change-point. Information about the hyperparameters and their respective prior scales can be found in the appendix.

First of all, there is a noticeable difference in the performance of vanilla Fbprophet model, which does not include any extra regressors or parameter tuning, compared to the same model with Finnish holiday added. MAE drops by approximately 1 with the information of Finnish holiday incorporated. On the other hand, the weather extra regressors do not deliver as much of predictive performance as initial expectation. The reduction in MAE and MAPE from weather variables is marginal compared to holiday. In the 16 Fbprophet trained with weather regressors, the top 10 sets of variables that produced the best prediction for 2019 can be found in the appendix 1. The snow variable appears in the best 8 models, implying its significance in predicting the taxi revenue, and the other three variables, which is minimum, maximum temperature and rain do not share the same level of importance. This contradicts with Kamga et al's finding [3] that snow conditions do not affect the hourly revenue. However, the authors employed New York City dataset in 2013, whose distribution might differ from the current dataset in question. This merits further studies to explore the difference. However, in this paper, the impact of weather to the taxi revenue is not thoroughly investigated. This brief exploration only serves the purpose of fine-tuning the prediction model.

However, the biggest improvement comes from the manual addition of change-point. When adding the manual changepoint of 01.06.2018, which is the disruptive change coming from the taxi deregulation, which changed the trajectory of taxi revenue, the error rate of the model was massively improved, thanks to the model's capability to adapt its prediction to the change in trend. The improvement can also be observed from figure 11.

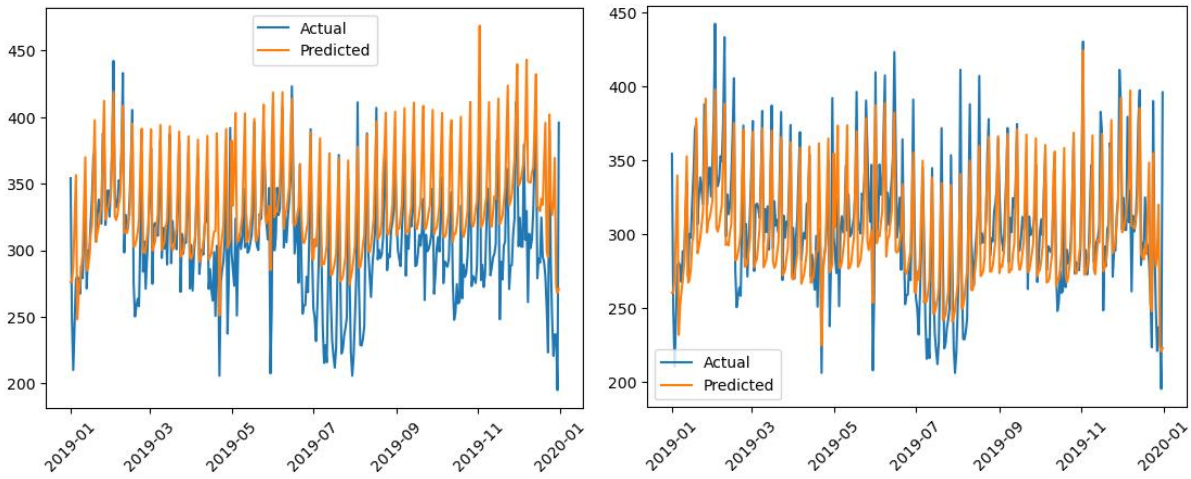

Figure 12: Prediction vs actual observations of daily taxi revenue per shift in 2019 without manual change point (left) and with manual change-point in 2019 (right)

From the findings in the experiments, the Fbprophet is tuned with increased prior scales for more important variables, such as the holidays and the snow independent variable. The model hyperparameters can be found in the appendix 2. This optimized Fbprophet model was able to obtain substantially improved performance, as the evaluation metrics can be seen from table 2. The prediction of the model and the actual data for 2019 is reported in the figure below. The model managed to obtain the seasonality relatively well. However, there was a disruptive change in the taxi business, in the form of the industry deregulation in July 2018, which introduced ride-hailing companies as fierce competitors to tradition taxi companies. Since then, taxi revenue in the dataset has experienced a negative change in the overall trend. Without the adding of change-point, the model was not able to capture the downward trajectory in the taxi revenue. However, thanks to the change-point addition, the model was able to fit the data relatively well, in terms of trend and seasonality. The performance of these models can be compared from figure 12.

5.2 Model prediction for 2020

Using the same model with optimized parameters in the previous section, the model is re-trained, with the training set including 2019 data as well. The examine period of the training set is from 01.01.2007 to 31.12.2019, which contains 4726 daily observations. The examine period of the testing set is from 01.01.2020 to 21.11.2020, which contains 324 daily observations.

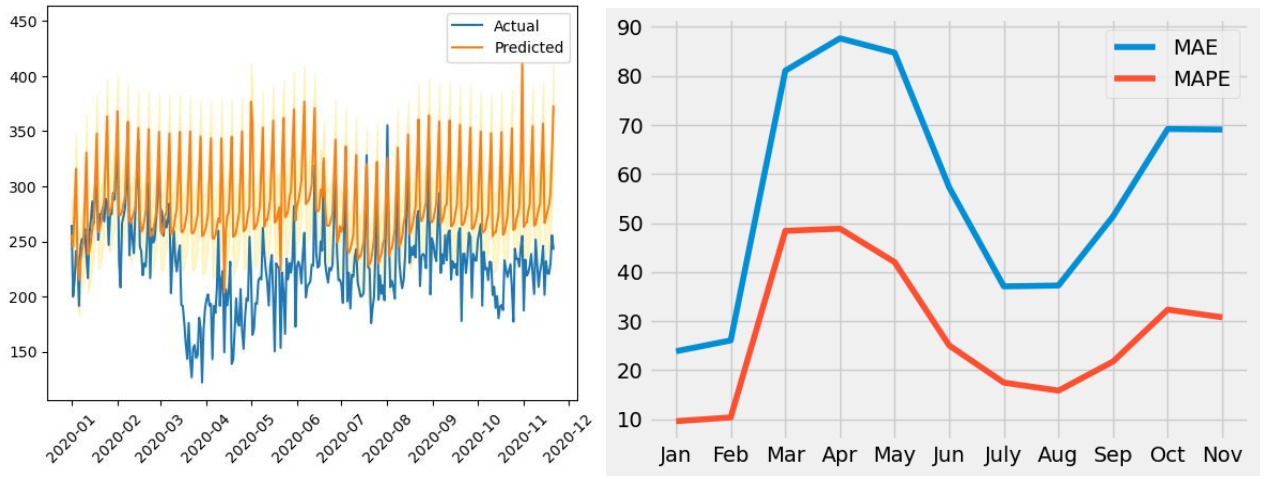

Figure 13: Prediction vs actual observations of taxi revenue in 2020 (left) and the evaluation metrics for each month of 2020 (right)

The figure on the left report the model prediction taxi revenue per shift, including the prediction value as well as the lower and upper bound of the prediction. The forecast is plotted against the actual data for 01.01.2020 until 21.12.2020. The figure on the right reports the evaluation metrics for 2020, MAE and MAPE. For the first two months January and February, the model performed relatively well in terms of MAE and MAPE. The evaluation metrics were at the same rate as MAE and MAPE of 2019's model, which means the forecast values from the model was performing within the expectation. However, when the first wave of COVID-19 pandemics hit around March 2020, the error rates MAE and MAPE skyrocketed, nearly four times larger than those of the previous months, and the error rates remained relatively high until May. When the total lockdown ended at the beginning of June, the taxi revenue started to stabilize, and MAE and MAPE of June, July and August reduced to an acceptable rate.

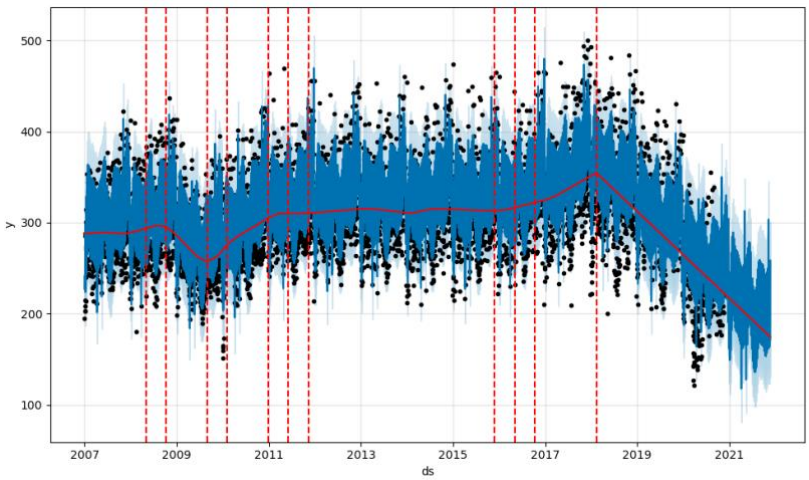

Figure 14: Forecast for 2021's taxi revenue per shift

Figure 14 shows the prediction of the Fbprophet model for the year 2021. The red lines represent the change-points in the trend. As can be seen from the graph, there was a downward trend from 2018. This can be explained by the impact of the taxi deregulation in 2018, which introduced more competition to the traditional taxi in terms of ride-sharing competition in Finland. Furthermore, COVID-19 pandemics has further reduced taxi revenue in 2020, which reinforced this negative trend. Thus, the model carries the negative trend to 2021 predictions. However, the trajectory of taxi revenue may very likely change, with the new COVID-19 vaccines being at advanced stage, where FDA panel recommended approvals for Pfizer's vaccines just on December 10th, 2020 [4], COVID-19 vaccines might be rolling out relatively soon. This is an excellent news for the whole economy as well as the Finnish taxi industry, since this will very likely boost the confidence of consumers and help recovering the Finnish economy as well as other industries, including the taxi industry. On the other hand, according to dialogues with experts from Exim.fi, Finland taxi industry can see changes in taxi regulations, similar to the deregulation in 2018. Therefore, the prediction from the Prophet model should only be used a reference for future work, instead of business decision making.

6 Conclusions and discussions

With a provided dataset of taxi shifts from Trafore Oy, this study utilizes multiple data analysis and time series correlation techniques to explore the impact of COVID-19 pandemics to the taxi industry in Finland. COVID-19 pandemics was estimated using indicators such as Google Trend search interest for keyword “COVID”, as well as the daily new cases of COVID-19 in Finland. Regarding the taxi industry, taxi revenue per shifts and number of shifts are used as a performance indicator. The study found a strongly negative correlation between taxi revenue per shift and Google search interest. Furthermore, an even stronger cross correlation of lag 1 and 2 is found between the aforementioned variables.On the demand side, the taxi revenue per shift experienced a staggering drop of 70% within March. Since the drop in March, it has gradually increased, but failed to reach the rate of previous years. The impact of COVID-19 is the most apparent from March to June, and from September onwards, with March and April being the most affected months in terms of taxi revenue per shifts. With the obtained insights from the study, an attempt to predict the taxi revenue rate was made using Prophet. However, the upcoming trajectory of the taxi industry is uncertain, with disruptive changes on the horizon, such as the end of COVID-19 pandemics and changes to the taxi regulation.

This study contains several limitations. The geographical distribution of taxi companies in the provided dataset are heavily skewed towards the Uusimaa area, with nearly 20% of the total number of registered customers in the dataset are from either Helsinki, Vantaa, or Espoo. This might reduce the generalization level of this paper to the whole country. On the other hand, the amount of data in research may not represent a whole picture of the whole Finnish industry in general. In 2019, the total accumulated amount of revenue from the dataset provided by Exim.fi is 25 million Euros, whereas the total revenue of taxis in Finland is 1.1 billion Euros according to Finnish Taxi Association (2020). Thus, the provided dataset only accounts for approximately 2.27%. This might not be representative of the taxi industry of Finland. Furthermore, the weather data was collected from the Helsinki base station, which could explain why the weather extra regressors was not able to provide the predictive power as expected.

Appendix

1. Top 10 Fbprophet models and their respective set of weather regressors. “False” indicates that the variable is not included in the model, and “True” indicates that the variable is included.

```
{'max_temp': False, 'min_temp': False, 'rain': True, 'snow': True}
{'max_temp': False, 'min_temp': False, 'rain': False, 'snow': True}
{'max_temp': True, 'min_temp': True, 'rain': False, 'snow': True}
{'max_temp': False, 'min_temp': True, 'rain': False, 'snow': True}
{'max_temp': True, 'min_temp': False, 'rain': True, 'snow': True}
{'max_temp': True, 'min_temp': True, 'rain': True, 'snow': True}
{'max_temp': True, 'min_temp': False, 'rain': False, 'snow': True}
{'max_temp': False, 'min_temp': True, 'rain': True, 'snow': True}
{'max_temp': False, 'min_temp': False, 'rain': False, 'snow': False}
{'max_temp': False, 'min_temp': False, 'rain': True, 'snow': False}
```

2. Hyperparameters of the final Fbprophet prediction model. The holiday and change point prior was chosen based on experimentation and choosing the performing model on 2019 subset of data.

```
train_model = Prophet(holidays_prior_scale=0.25,
                      changepoint_prior_scale=0.5,
                      n_changepoints=50,
                      changepoints=['2018-06-01'],
                      weekly_seasonality=True)
train_model.add_country_holidays(country_name='FI')
train_model.add_regressor('min_temp', prior_scale=0.05)
train_model.add_regressor('max_temp', prior_scale=0.05)
train_model.add_regressor('rain', prior_scale=0.05)
train_model.add_regressor('snow', prior_scale=0.3)
```

References

[1] A. Svynarenko and M. Perkiö, “Taxi platforms respond to COVID-19”, April 2020. [Online] Available at: https://alusta.uta.fi/2020/04/02/taxi-platforms-respond-to-covid-19/. [Accessed Sep 27, 2020].

[2] Airlines for American, “Impact of Covid 19”, 2020. [Online] Available at: https://www.airlines.org/dataset/impact-of-covid19-data-updates/#. [Accessed Sep 27, 2020].

[3] C.N. Kamga, M.A. Yazici, A. Singhai, “Hailing in the Rain: Temporal and Weather-Related Variations in Taxi Ridership and Taxi Demand-Supply Equilibrium”, Jan 2013. [Online] Available at: https://www.researchgate.net/publication/255982467_Hailing_in_the_Rain_Temporal_and_Weather-Related_Variations_in_Taxi_Ridership_and_Taxi_Demand-Supply_Equilibrium. [Accessed Oct 1, 2020].

[4] CNBC, “FDA panel recommends approval of Pfizer’s Covid vaccine for emergency use”, 2020. [Online] Available at: https://www.cnbc.com/2020/12/10/pfizer-covid-vaccine-fda-panel-recommends-approval-for-emergency-use.html. [Accessed Oct 1, 2020].

[5] D. Tahlyan and H.S. Mahmassani, “Chicago Mobility under COVID-19”, 2020. [Online] Available at: https://www.transportation.northwestern.edu/news-events/articles/2020/chicago-mobility-under-covid-19.html. [Accessed Oct 1, 2020].

[6] Finlex, ”24.5.2017/320: Laki liikenteen palveluista”. [Online] Available at: https://www.finlex.fi/fi/laki/ajantasa/2017/20170320. [Accessed Oct 1, 2020].

[7] Finnish Meteorological Institute, “Download observations”, 2020. [Online] Available at: https://en.ilmatieteenlaitos.fi/download-observations. [Accessed Oct 1, 2020].

[8] Google, “How Trends data is adjusted”, Nov 8, 2017. [Online] Available at: https://support.google.com/trends/answer/4365533?hl=en. [Accessed Oct 1, 2020].

[9] H. Ale-Ahmad and H.S. Mahmassani, “Impact of COVID-19 on Taxi Operation in Chicago” 2020. [Online] Available at: https://www.transportation.northwestern.edu/news-events/articles/2020/taxi-operations-during-covid-19.html. [Accessed Oct 1, 2020].

[10] H. Metsäranta, J. Tervonen and V. Jaakola, ”Taksiliikenteen hintaseuranta ja enimmäishinnan laskentaperusteet”, 2018. [Online] Available at: https://arkisto.trafi.fi/filebank/a/1530527390/24532c80c9f5f9b02070268c62fe3c46/31128-Trafin_julkaisuja_15_2018_Taksiliikenteen_hintaseuranta_ja_enimmaishinnan_laskentaperusteet.pdf [Accessed Oct 1, 2020].

[11] Helsinki Times, “Benefits of taxi industry reform are few and far between, shows survey”, March 2019. [Online] Available at: https://www.helsinkitimes.fi/finland/finland-news/domestic/16281-benefits-of-taxi-industry-reform-are-few-and-far-between-shows-survey.html. [Accessed Sep 27, 2020].

[12] K. Väyrynen, “Ride-Hailing App Strategies of Finnish Taxi Dispatch Organizations”, 2020. [Online] Available at: https://press.um.si/index.php/ump/catalog/view/483/587/963-1. [Accessed Sep 27, 2020].

[13] P. Sandholm, “Consequences of the deregulation of taxi service”, 2019. [Online] Available at: https://www.theseus.fi/handle/10024/266806. [Accessed Oct 1, 2020].

[14] RDRR, “bpr.function: BPR cost and objective functions”, May 2, 2019. [Online] Available at: https://rdrr.io/rforge/travelr/man/bpr.function.html. [Accessed Oct 1, 2020].

[15] Shenzhen Government Online, “Overview of Shenzhen”, 2017. [Online] Available at: https://web.archive.org/web/20170525115028/http://english.sz.gov.cn/gi/. [Accessed Oct 1, 2020].

[16] Suomen taksiliitto, “Coronavirus and school closure overthrew taxi operators”, March 16, 2020. [Online] Available at: https://www.taksiliitto.fi/koronavirus-ja-koulujen-sulkeminen-kaataa-taksiyrittajat/. [Accessed Sep 24, 2020].

[17] Suomen Taksiliitto, “Information about Finnish taxi sector”, June 2020. [Online] Available at: https://www.taksiliitto.fi/en/home-2/. [Accessed Sep 27, 2020].

[18] Taksi Rekisteri, “Information about Taxi Companies”, 2020. [Online] Available at: https://www.taksirekisteri.fi/stlextranet/otetietohaku.aspx. [Accessed Sep 27, 2020].

[19] THL, “Confirmed Corona cases in Finland (COVID-19)”, 2020. [Online] Available at: https://thl.fi/en/web/zfi-en/statistics/statistical-databases/open-data/confirmed-corona-cases-in-finland-covid-19-. [Accessed Oct 1, 2020].

[20] Traficom,“Liikennelupatilastot”, 2019. [Online] Available at: https://www.traficom.fi/fi/liikennelupatilastot. [Accessed Oct 1, 2020].

[21] UNWTO, “INTERNATIONAL TOURIST NUMBERS COULD FALL 60-80% IN 2020, UNWTO REPORTS”, 2020. [Online] Available at: https://www.unwto.org/news/covid-19-international-tourist-numbers-could-fall-60-80-in-2020. [Accessed Sep 27, 2020]..

[22] X. Qian and S.V. Ukkusuri, “Modeling the spread of infectious disease in urban areas with travel contagion”, 2020. [Online] Available at: http://arxiv.org/abs/2005.04583, arXiv:2005.04583. [Accessed Oct 1, 2020].

[23] Y. Hu, W. Barbour, S. Samaranayake, and D. Work, “Impacts of covid-19 mode shift on road traffic”, 2020. [Online] Available at: https://arxiv.org/abs/2005.01610. [Accessed Oct 1, 2020].

[24] Y. Nie. “How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China”, June 2017. [Online] Available at: https://www.sciencedirect.com/science/article/pii/S0968090X17301018. [Accessed Sep 27, 2020].

[25] Yle, "Passengers from Bucharest tested on arrival at Helsinki Airport”, August 8, 2020. [Online] Available at: https://yle.fi/uutiset/osasto/news/passengers_from_bucharest_tested_on_arrival_at_helsinki_airport/11484992. [Accessed Oct 1, 2020].

[26] Yle, “Finnair plans further flight reductions in October”, 2020. [Online] Available at: https://yle.fi/uutiset/osasto/news/finnair_plans_further_flight_reductions_in_october/11535380. [Accessed Sep 27, 2020].

[27] Yle, “Health officials: As infections rise, mask recommendation likely, possibly on regional basis”, August 6, 2020. [Online] Available at:

https://yle.fi/uutiset/osasto/news/health_officials_as_infections_rise_mask_recommendation_likely_possibly_on_regional_basis/11481643. [Accessed Oct 1, 2020].

[28] Yle, "Paper: Covid-19 cases confirmed among passengers from Skopje", August 10, 2020. [Online] Available at: https://yle.fi/uutiset/osasto/news/paper_covid-19_cases_confirmed_among_passengers_from_skopje/11487038. [Accessed Oct 1, 2020].

[29] Yle, " More events cancelled due to rise in infections", August 6, 2020. [Online] Available at: https://yle.fi/uutiset/osasto/news/more_events_cancelled_due_to_rise_in_infections/11485731. [Accessed Oct 1, 2020].

[30] N. Lomas, "Uber relaunches a licensed service in Finland after taxi law deregulation", July 4, 2018. [Online] Available at: https://techcrunch.com/2018/07/04/uber-relaunches-a-licensed-service-in-finland-after-taxi-law-deregulation/. [Accessed Oct 1, 2020].

[31] Yle, "Uber drivers lose test case in Helsinki Court", Sep 21, 2016. [Online] Available at: https://yle.fi/uutiset/osasto/news/uber_drivers_lose_test_case_in_helsinki_court/9181897. [Accessed Oct 1, 2020].

[32] Statista, "Ride-Hailing and Taxi in Finland", 2020. [Online] Available at: https://www.statista.com/outlook/368/135/ride-hailing-taxi/finland. [Accessed Oct 1, 2020].

[33] H. Zheng, K. Zhang and M. Nie, "The Fall and Rise of the Taxi Industry in the COVID-19 Pandemic: A Case Study", August 16, 2020. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3674241. [Accessed Oct 1, 2020].

[34] Facebook, "Forecasting at scale.", 2020. [Online] Available at: https://facebook.github.io/prophet/. [Accessed Oct 1, 2020].