esprit
Se former autrement

DATA SCIENCE PROJECT: SALES AND ITEMS PREDICTIVE SYSTEM

A WORK BY

JILENI ZEHANI

MITAKUS

Table of Contents

Table of Figures

Introduction

As it becomes ever easier and cheaper to collect and store vast amounts of data, all fields of human endeavor are transforming into data-intensive fields. No matter what career path a student selects, he or she will need to be adept at manipulating, analyzing, and interpreting data at scales never required previously. As William Buxton describes, technology has advanced to the point where the limiting factor currently is human capability.

Today’s “big data” requires more processing in order to extract knowledge from data sets that are often noisy, in inconvenient formats, and, simply, too large to useful unless somehow presented on a more “human” scale.

Companies need to use data to run and grow their everyday business. The fundamental goal of data science is to help companies make quicker and better decisions, which can take them to the top of their market, or at least – especially in the toughest red oceans – be a matter of long-term survival.

Chapter 1: Context Analysis

1. Global Context

Mitakus' job is to help chefs and owners of canteens, restaurants and other gastronomy businesses with actionable analytics and business insights.

Our mission during this project is to help owners and chefs in gastronomy to have a more profitable business by cutting costs, especially in the form of food waste, and maximizing revenue.

2. Project Context

A canteen, like any other business, may face different types of challenges in a rapidly changing and evolving world, with demanding gourmets, especially if it aspires to improve the quality of food and services, increase profitability and reduce costs in a short time while preserving the environment and avoiding wastage. These challenges are not very different from those of an ordinary restaurant, and they're mainly related to inventory management and menu engineering problems, caused by the inability of accurately predicting the future demand, sales and trends. Great amounts of food are wasted every year due to the overestimation of the sale of certain dishes, and a lot of money would've been gained if the shortage of food and ingredients had been avoided. The offered menu can also be improved and edited to insure a continuous and increasing satisfaction of customers, items can be added and others can be taken down either temporarily or permanently.

Chapter 2: Business Objectives and Data Science Goals

1. Business Understanding

Canteens are a renowned food service business that is found within every establishment, often a larger institution, catering to the clientele of that institution. For example, schools, colleges and many other places such as office buildings, factories and hospitals.

A canteen provides the best conditions for earning a profit by assuring:

- A Captive Market: A canteen provides the establishment with another source of revenue and a captive market for the operator.
- Lower Marketing Expenses: With a captive market, the canteen operator does not have to spend much on marketing and promotional activities.
- Lower Overhead Costs: The rental rates can be exorbitant. In some places, rent is calculated at 34% of gross sales. So if your food cost is 50%, what are your chances of making a profit once you have paid off wages, utilities, and benefits?

For a canteen with a serving capacity of some 200 people, an initial investment of around $2 million to $3.5 million is required. About 70% of that amount would go to acquiring equipment, and the rest could serve as the initial working capital.

For an Office Canteen, it gives building tenants and visitors an affordable option to eat hot home-cooked meals without having to worry about time, traffic and unpredictable weather.

Over the last ten years, the office canteen has evolved into exquisite food courts. There are more food varieties, and the interiors are elegant.

While you do get good foot traffic here, business on weekends screeches to a near halt.

A canteen is said to have a captured market, whether it is located in a school, office building or factory. This does not make running the business any easier. It just presents a particular set of challenges depending on the market one chooses to serve.

Since a canteen has a captured market, it needs a menu that offers sufficient variety and for the owners to regularly introduce new items to the menu.

There are a lot of items to track when operating a canteen. These include cooked dishes down to the last cooking ingredient, cooking and eating utensils, and kitchen equipment.

“You serve the same people every day and they should not get tired of your offerings,” says Marie Paz Pineda, a renowned food concessionaire.

2. Business Objectives

For every business, making its products or services better is the ultimate goal of a data science project.

Well-chosen goals and objectives point a new business in the right direction and keep an established company on the right track, therefore, we opted for the following business objectives:

- Sustainable growth
- Menu engineering
- Customer attraction and retention
- Maximizing business profitability
- Marketing based on sale forecasts

3. CRISP-DM Methodology

In order to ensure quality in our data science project we should make sure that we are enforcing a standard methodology, that’s why we opted for the CRISP-DM methodology since it provides strong guidance for even the most advanced of today’s data science activities. CRISP-DM stands for cross-industry process for data mining.

The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. We do not claim any ownership over it. We did not invent it. We are however evangelists of its powerful practicality, its flexibility and its usefulness when using analytics to solve thorny business issues. It is the golden thread that runs through almost every client engagement.

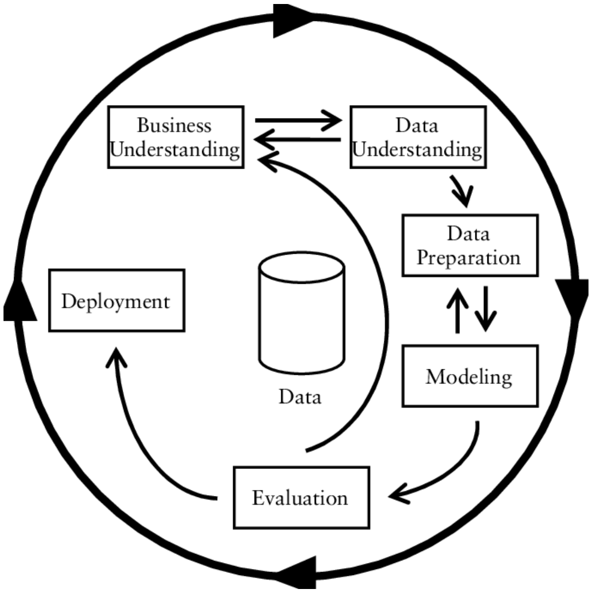


Figure 1: The CRISP-DM Model

4. Data Science Goals

In order to achieve our business objective, we have to set the data science goals which consists of:

- Predicting revenue and profit using time series and regression
- Evaluating items in terms of quantity using process mining
- Item clustering (breakfast/lunch /snacks)
- Evaluating items in terms of revenue using process mining
- Predicting high/low demand periods using time series and regression

5. Conclusion

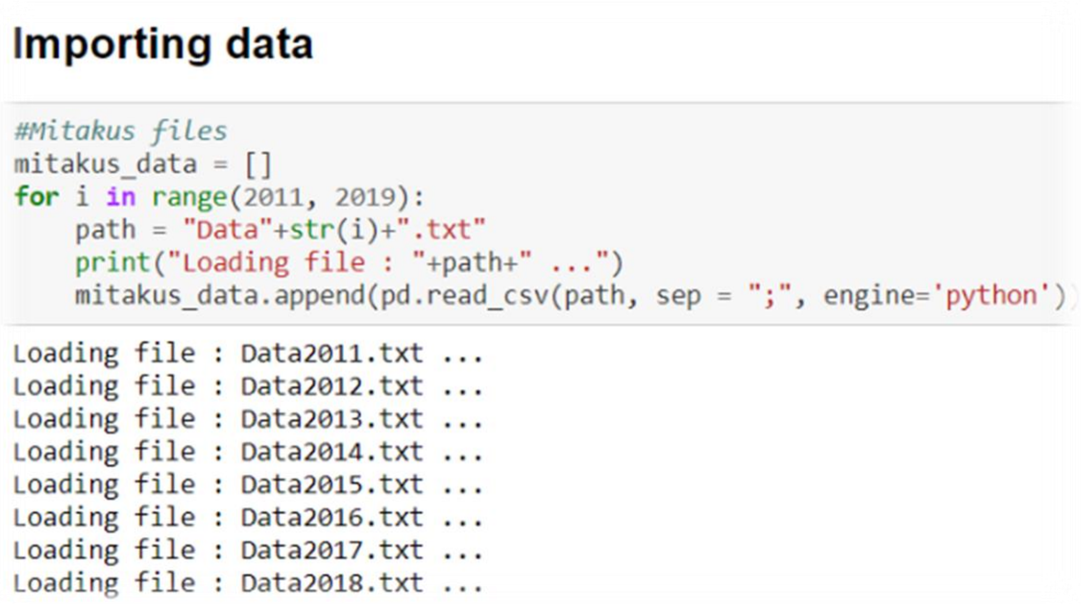
A somewhat sophisticated forecasting system based on data collected from canteens can be a great tool for owners to better understand the complex eating behavior and food choice of their clients, the different and unexpected trends that may or repeatedly occur and know exactly how to overcome each and every one of the challenges their business may face with the most optimized approach possible, to increase profit and grow sustainably while maintaining a high quality of service and high competitiveness while avoiding as much unnecessary squandering as possible.

Chapter 3: Data Understanding and Preparation

1. Data Collection

Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis based on collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

In our case the data is provided by Mitakus, it is mainly the sales records of two canteens located in Lehel and Giesing. The records provided are from



Importing data

```
#Mitakus files
mitakus_data = []
for i in range(2011, 2019):
    path = "Data"+str(i)+".txt"
    print("Loading file : "+path+" ...")
    mitakus_data.append(pd.read_csv(path, sep = ";", engine='python'))
```

```
Loading file : Data2011.txt ...
Loading file : Data2012.txt ...
Loading file : Data2013.txt ...
Loading file : Data2014.txt ...
Loading file : Data2015.txt ...
Loading file : Data2016.txt ...
Loading file : Data2017.txt ...
Loading file : Data2018.txt ...
```

Figure 2: Data Importing

2011 to 2018. All the records are merged in a single dataset.

2. Data Cleaning

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the coarse data.

In this phase, we did the following steps:

- Verifying missing data
- Spot and drop duplicates
- Handling outliers
- Summarizing data

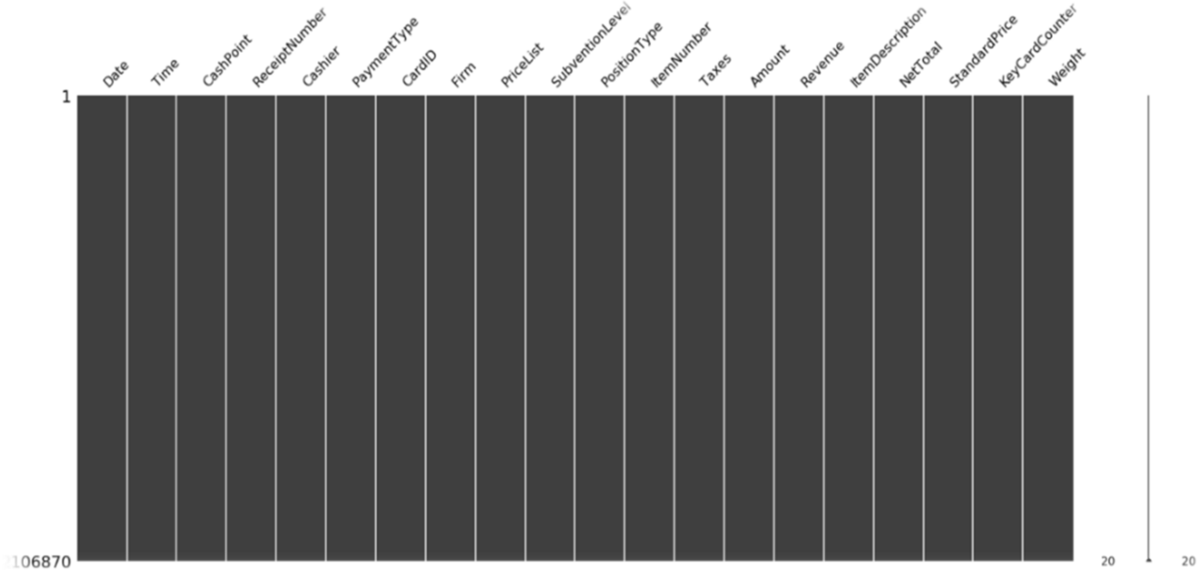


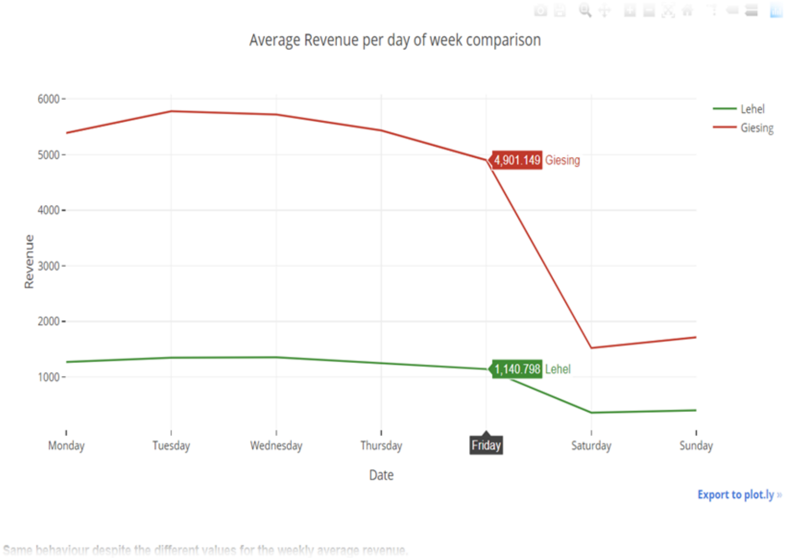
Figure 3: Missing Values Checking

3. Data Visualization

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.

For revenue and items:

- Overall Revenue
- Resampling by year
- Resampling by month
- Resampling by week
- Resampling by day
- Top items



Same behaviour despite the different values for the weekly average revenue.

Figure 4: Average Revenue per Day of Week Comparison

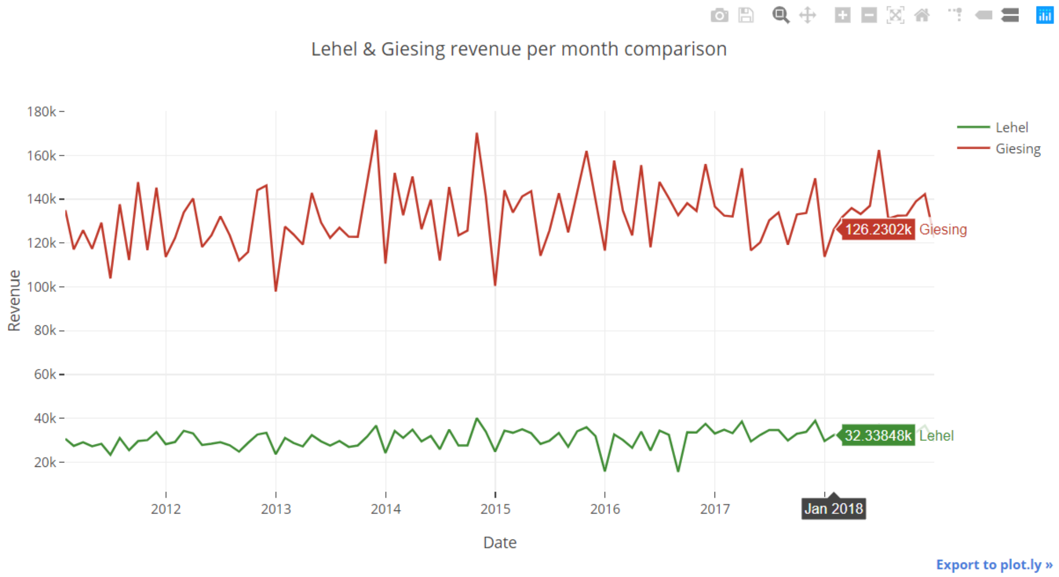


Figure 5: Lehel and Giesing Revenue per Month Comparison

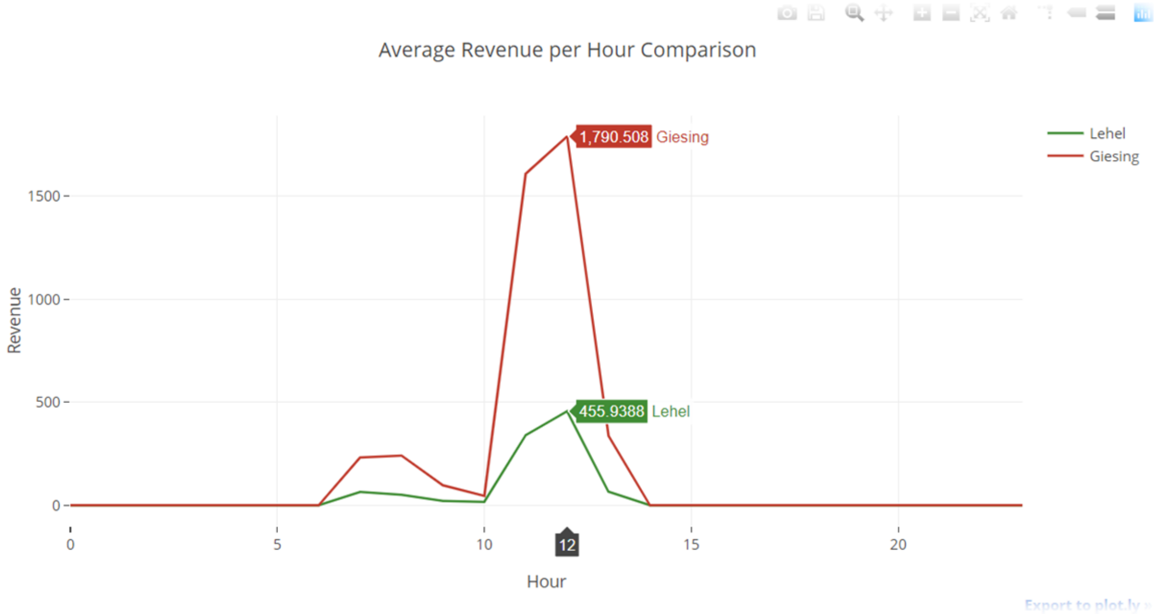


Figure 6: Average Revenue per Hour Comparison

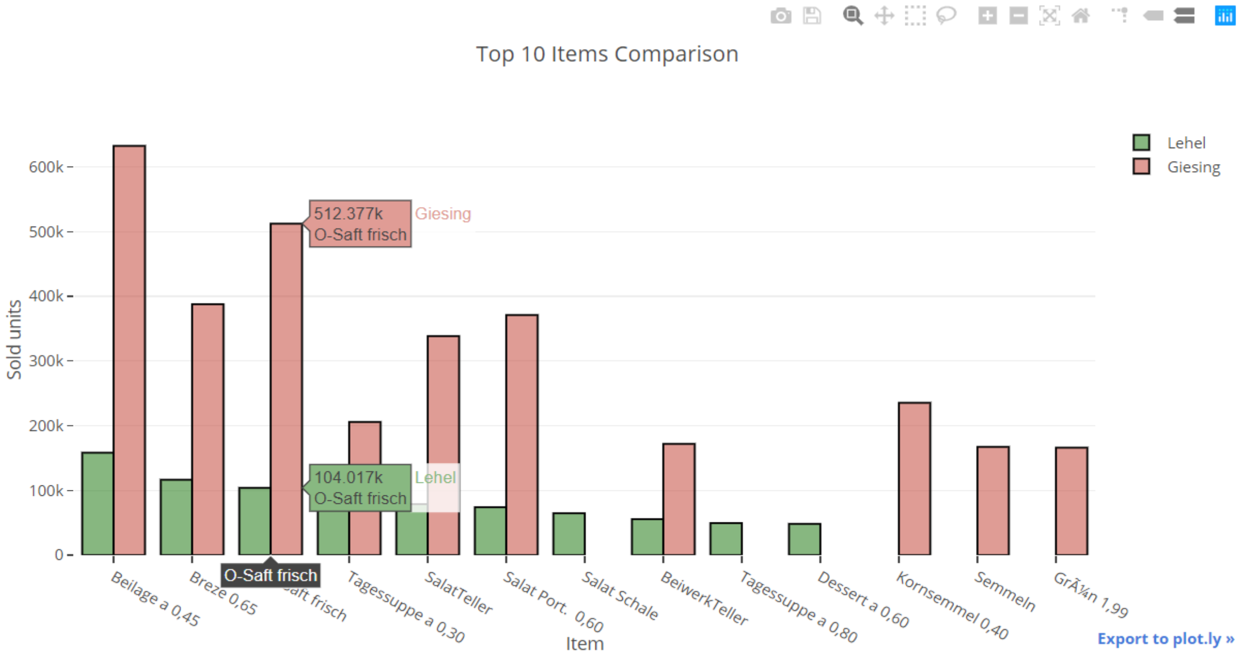


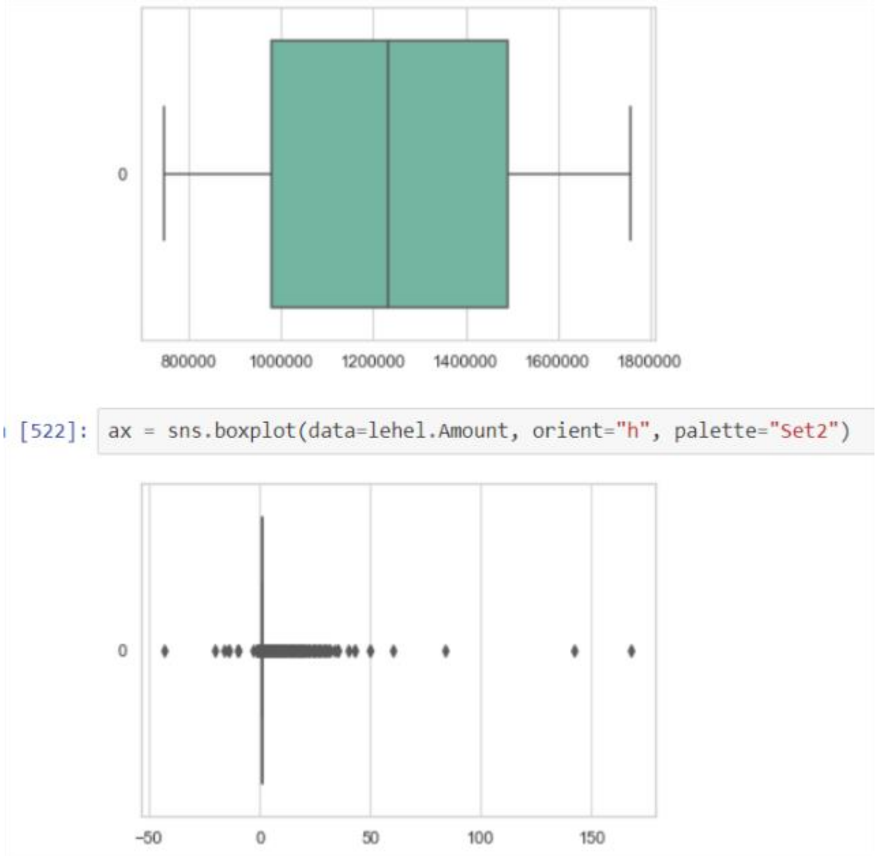
Figure 7: Top 10 Items Comparison

4. Feature Selection

Getting data ready for modelling, one of the main reasons that causes machine-learning models to overfit is because of having redundancy in our data, which makes the model to be too complex for the given training data and unable to generalize well on unseen data.

Univariate Feature Selection This technique involves more of a manual kind of work. Visiting every feature and checking its importance with the target.

Also checking the variance, the thumb rule here is set a threshold value (say a feature with 0 variance means it has the same value for every sample so such a feature would not bring any predictive power to the model) remove feature accordingly.



```
[522]: ax = sns.boxplot(data=lehel.Amount, orient="h", palette="Set2")
```

Figure 8: Boxplots of the Variables Receipt Number and Amount

Chapter 4: Modelling and Evaluation

1. Proposed Models

The main goal of our project is forecasting the revenue and the profit of the canteen, predicting high/low demand periods and finally evaluating items in terms of revenue/quantity using process mining, which can be achieved with Regression models.

Regression Models consists of a set of machine learning methods that allow us to predict a continuous outcome variable, the revenue and the profit in our case, based on the value of one or multiple predictor variables (in our case based on the time series).

In order to achieve best results, we used the following regression models:

- Gradient Boosting Machine

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

- Random Forrest Regressor

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

- Catboost Regressor

Catboost Regressor is a light gradient boosting based on decision trees library.

- XGBoost (eXtreme Gradient Boosting)

XGBoost is an algorithm that has recently been dominating applied machine learning, it consists of an implementation of gradient boosted decision tree, designed for speed and performance.

- LSTM Neural Network

A powerful type of neural network designed to handle sequence dependence is called recurrent neural networks. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained.

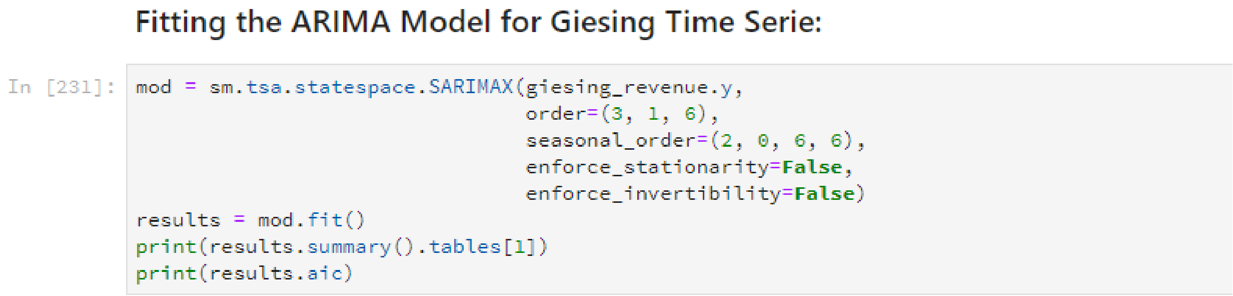
- ARIMA (Auto Regressive Integrated Moving Average)

ARIMA is a class of model that captures a suite of different standard temporal structures in time series data, it provides a simple yet powerful method for making skillful time series forecasts.

2. **Building Models**

Building models to address specific business goals is the heart of the data-science profession.

For this task, we used “scikit-learn” (a machine learning library in Python) for the Random Forrest Regressor and XGBoost Model. We also used the lightgbm as well as catboost open source libraries. For the ARIMA models we used “statsmodels” python module. We used grid searching to optimize and tune our models.



Fitting the ARIMA Model for Giesing Time Serie:

```
In [231]: mod = sm.tsa.statespace.SARIMAX(giesing_revenue.y,
                                          order=(3, 1, 6),
                                          seasonal_order=(2, 0, 6, 6),
                                          enforce_stationarity=False,
                                          enforce_invertibility=False)
          results = mod.fit()
          print(results.summary().tables[1])
          print(results.aic)
```

Figure 9: Fitting the ARIMA Model

- **Train-Test Splitting:**

We often split our data into a train and a test set, the training set used to prepare the model and the test set used to evaluate it. We may even use k-fold cross validation that repeats this process by systematically splitting the data into k groups, each given a chance to be a held out model.

These methods cannot be directly used with time series data. Instead, we must split data up and respect the temporal order in which values were observed.

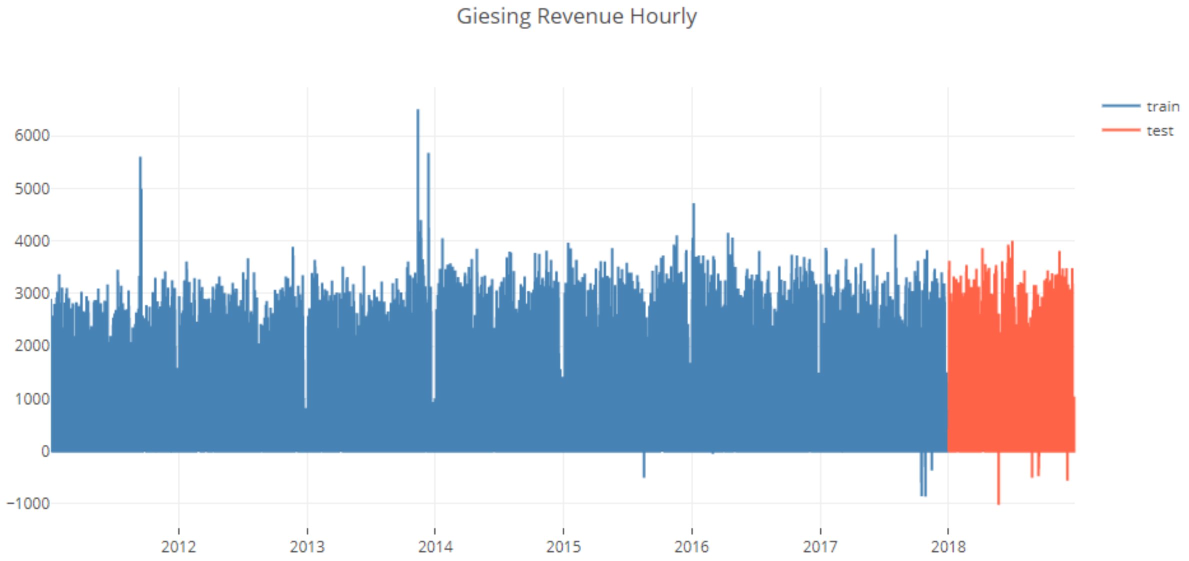


Figure 10: Giesing's Data Split into a Test and a Train Set

3. Model Evaluation

Model evaluation leads a Data Scientist in the right direction to select or tune an appropriate model.

Validation and Evaluation of a Data Science Model provides more color to our hypothesis and helps evaluate different models that would provide better results against our data. These are the metrics used in order to evaluate our models.

- Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- Root Mean Squared Error (RMSE):

RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- R2 Score (Coefficient of Determination)

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%.

$$R^2 \equiv 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

4. Forecasting Results

For each of our models, we did some visualization to compare our predictions and forecasts with real data.

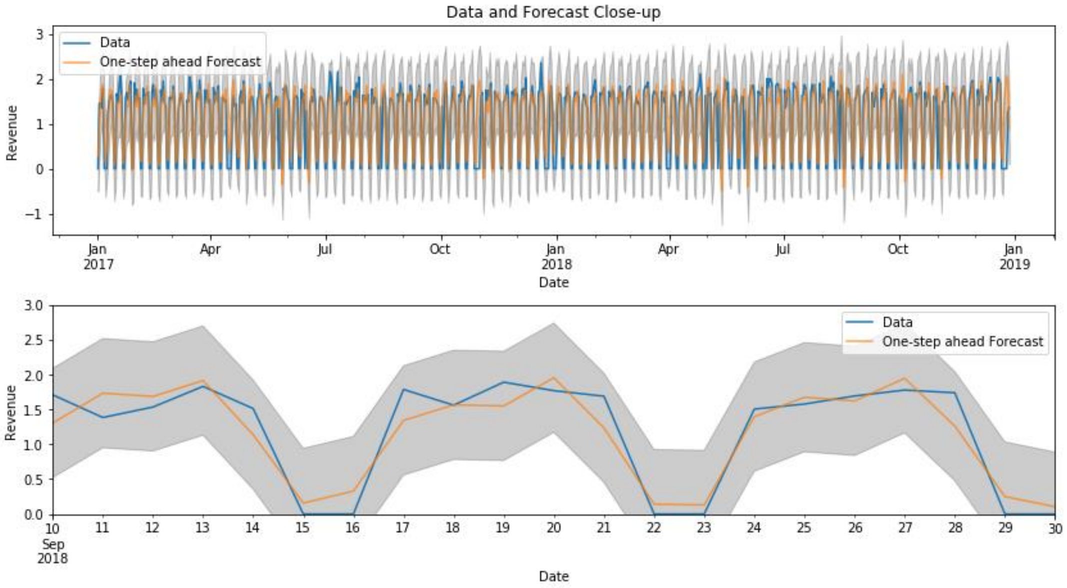


Figure 11: Forecast with ARIMA

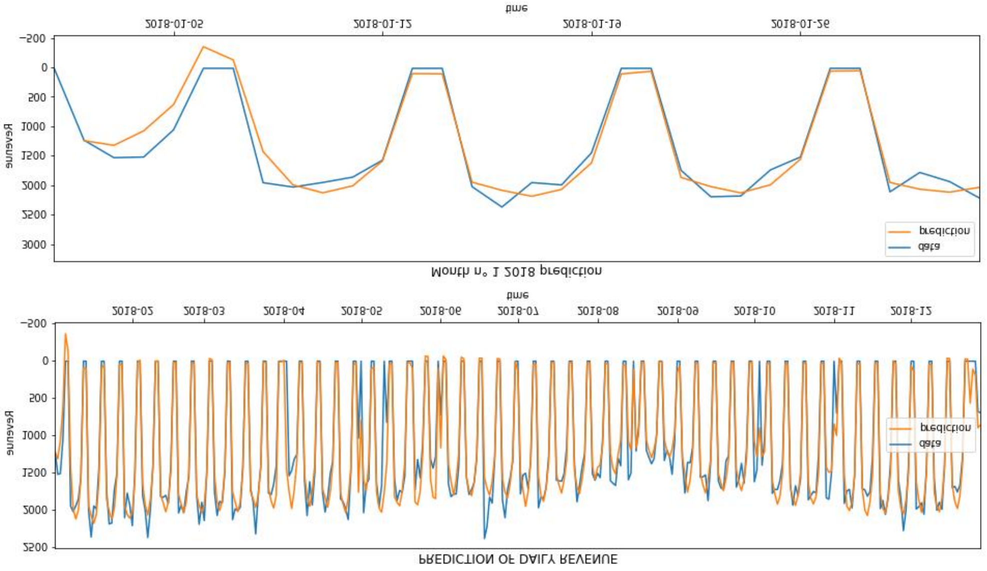


Figure 12: Prediction with LightGBM

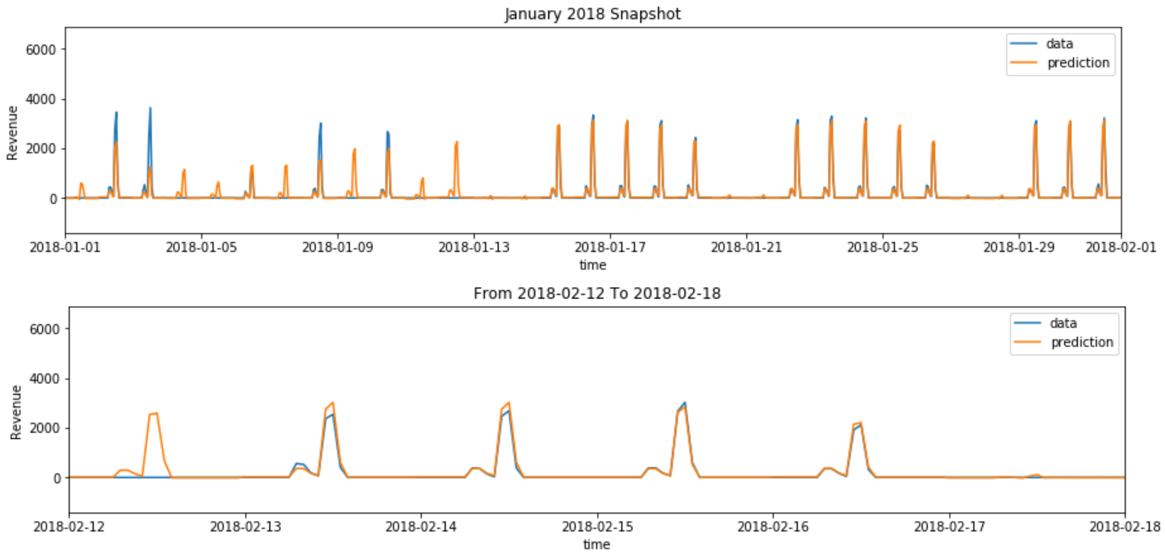


Figure 13: Prediction with XGBoost

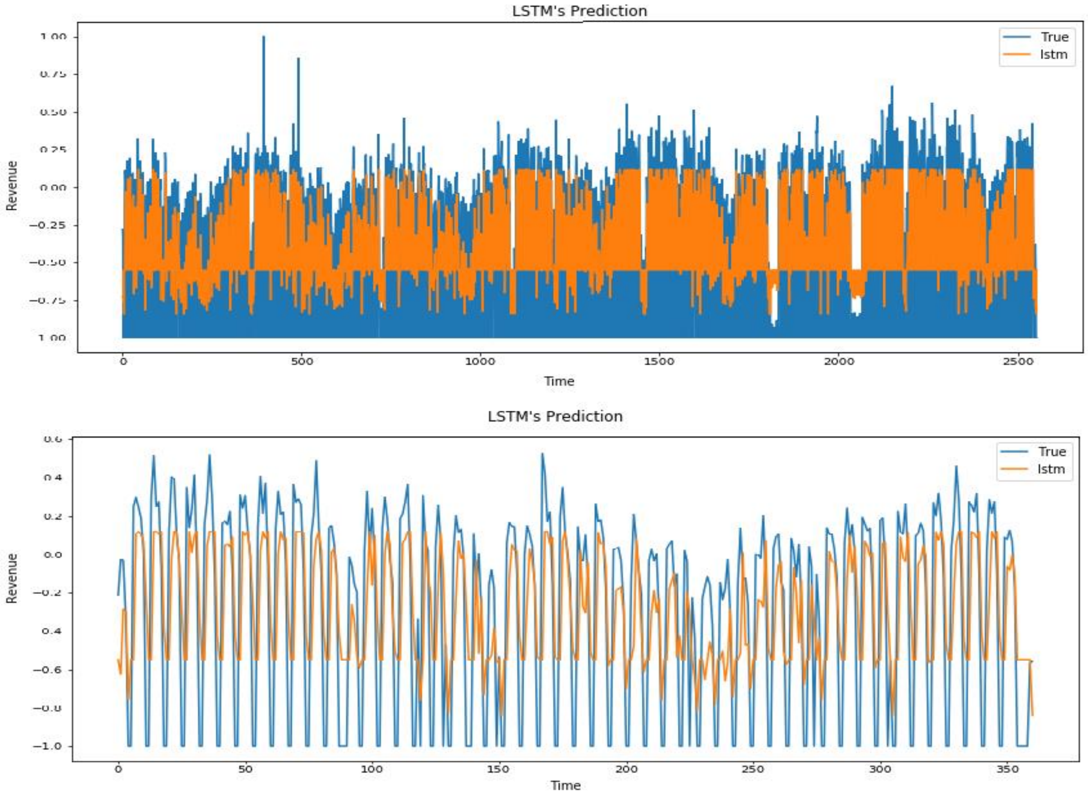


Figure 14: Prediction with XGBoost LSTM

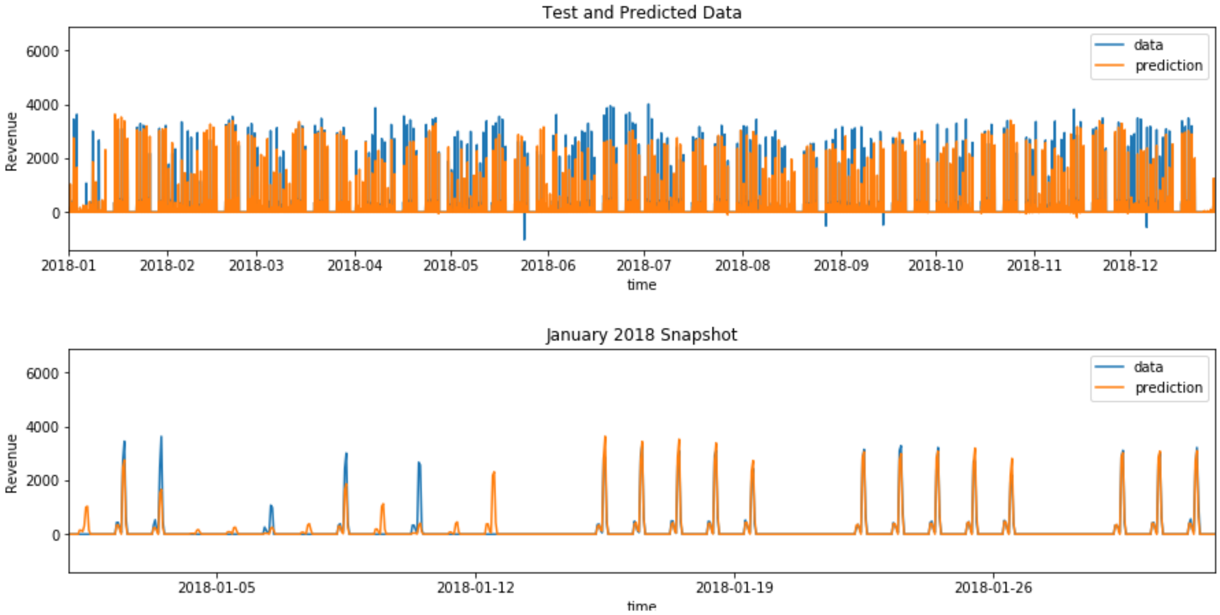
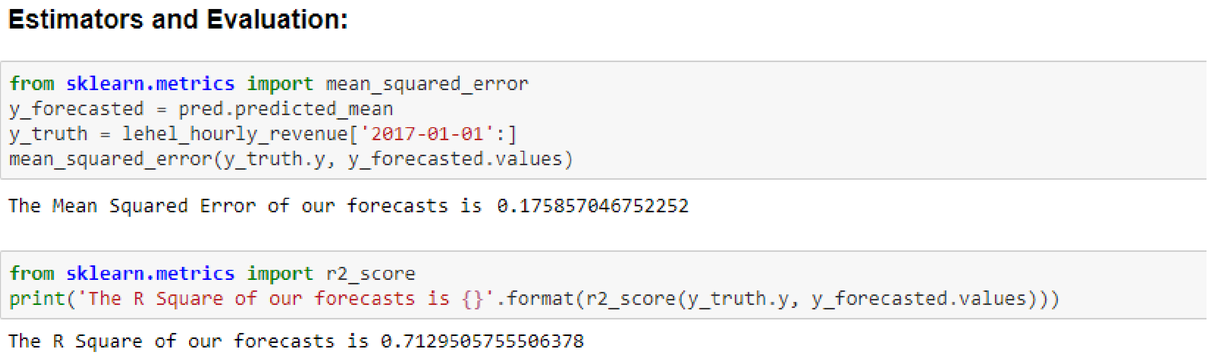


Figure 15: Prediction with Random Forest

We applied some of the measures mentioned earlier to evaluate our models, in figures below we have the evaluation for each of the models:



Estimators and Evaluation:

```
from sklearn.metrics import mean_squared_error
y_forecasted = pred.predicted_mean
y_truth = lehel_hourly_revenue['2017-01-01':]
mean_squared_error(y_truth.y, y_forecasted.values)
```

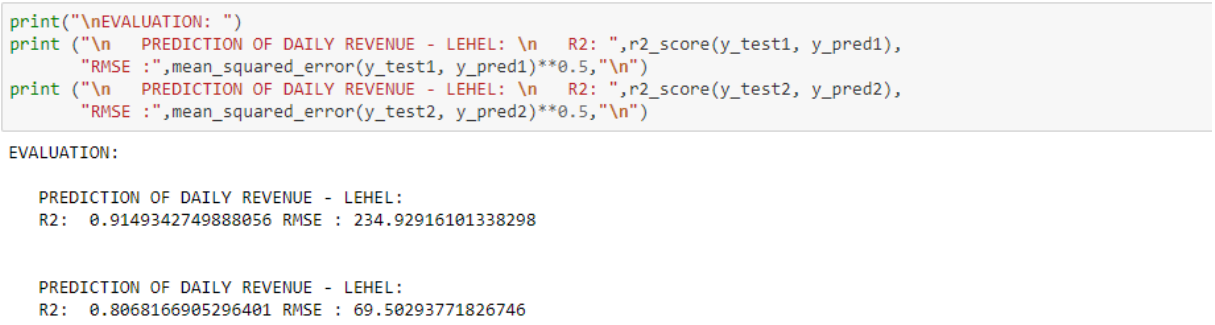
```
The Mean Squared Error of our forecasts is 0.175857046752252
```

```
from sklearn.metrics import r2_score
print('The R Square of our forecasts is {}'.format(r2_score(y_truth.y, y_forecasted.values)))
```

```
The R Square of our forecasts is 0.7129505755506378
```

Figure 16: ARIMA Evaluation

Evaluation:



```
print("\nEVALUATION: ")
print ("\n   PREDICTION OF DAILY REVENUE - LEHEL: \n   R2: ",r2_score(y_test1, y_pred1),
       "RMSE :",mean_squared_error(y_test1, y_pred1)**0.5,"\n")
print ("\n   PREDICTION OF DAILY REVENUE - LEHEL: \n   R2: ",r2_score(y_test2, y_pred2),
       "RMSE :",mean_squared_error(y_test2, y_pred2)**0.5,"\n")
```

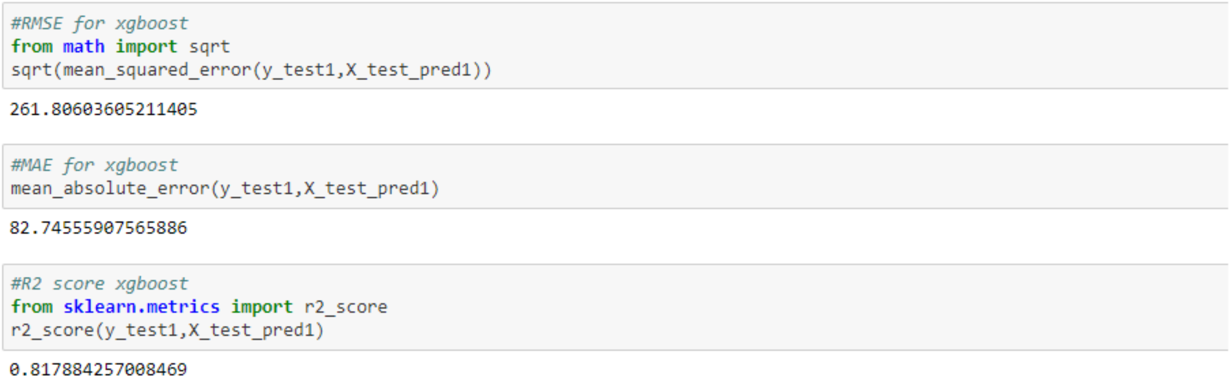
```
EVALUATION: 

   PREDICTION OF DAILY REVENUE - LEHEL: 
   R2:  0.9149342749888056 RMSE : 234.92916101338298 


   PREDICTION OF DAILY REVENUE - LEHEL: 
   R2:  0.8068166905296401 RMSE : 69.50293771826746 
```

Figure 17: LightGBM Evaluation

Measuring the error



```
#RMSE for xgboost
from math import sqrt
sqrt(mean_squared_error(y_test1,X_test_pred1))
```

```
261.80603605211405
```

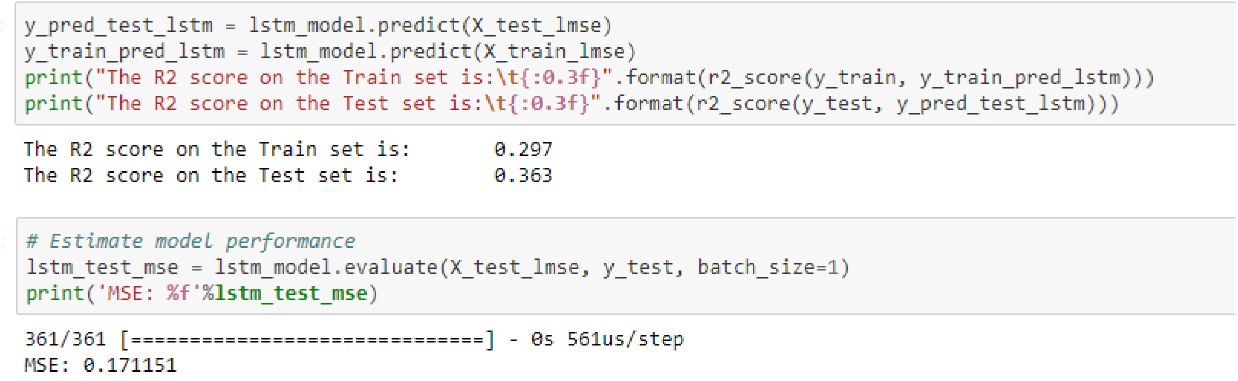
```
#MAE for xgboost
mean_absolute_error(y_test1,X_test_pred1)
```

```
82.74555907565886
```

```
#R2 score xgboost
from sklearn.metrics import r2_score
r2_score(y_test1,X_test_pred1)
```

```
0.817884257008469
```

Figure 18: XGBoost Evaluation



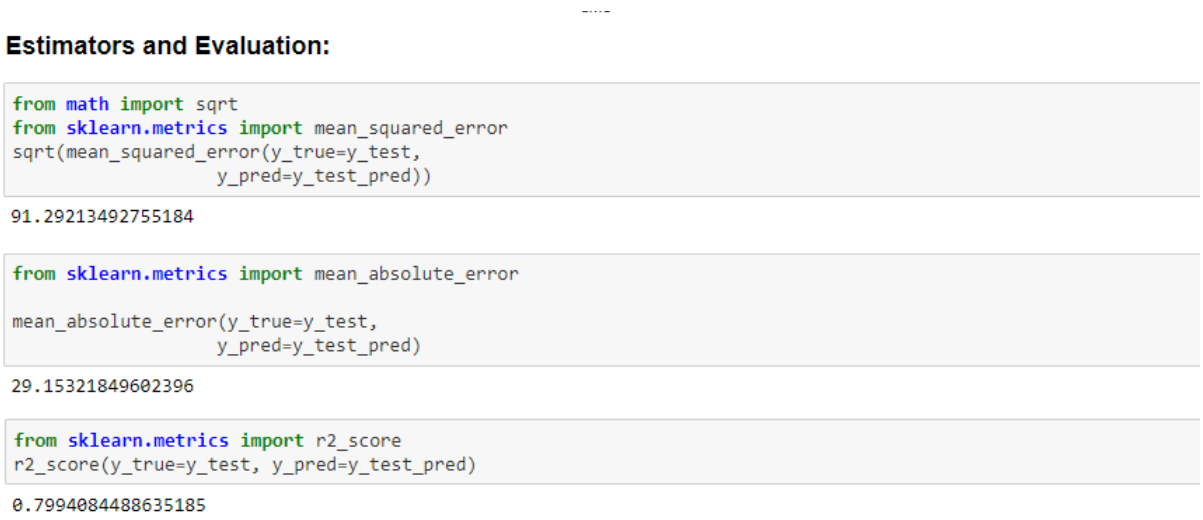
```
y_pred_test_lstm = lstm_model.predict(X_test_lmse)
y_train_pred_lstm = lstm_model.predict(X_train_lmse)
print("The R2 score on the Train set is:\t{:0.3f}".format(r2_score(y_train, y_train_pred_lstm)))
print("The R2 score on the Test set is:\t{:0.3f}".format(r2_score(y_test, y_pred_test_lstm)))
```

```
The R2 score on the Train set is:       0.297
The R2 score on the Test set is:        0.363
```

```
# Estimate model performance
lstm_test_mse = lstm_model.evaluate(X_test_lmse, y_test, batch_size=1)
print('MSE: %f'%lstm_test_mse)
```

```
361/361 [==============================] - 0s 561us/step
MSE: 0.171151
```

Figure 19: LSTM Evaluation



Estimators and Evaluation:

```
from math import sqrt
from sklearn.metrics import mean_squared_error
sqrt(mean_squared_error(y_true=y_test,
                   y_pred=y_test_pred))
```

```
91.29213492755184
```

```
from sklearn.metrics import mean_absolute_error

mean_absolute_error(y_true=y_test,
                   y_pred=y_test_pred)
```

```
29.15321849602396
```

```
from sklearn.metrics import r2_score
r2_score(y_true=y_test, y_pred=y_test_pred)
```

```
0.7994084488635185
```

Figure 20: Random Forest Evaluation

Chapter 5: Deployment

1. Introduction

Deploying models is the key to making them useful. The concept of deployment in data science refers to the application of a model for prediction using a new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process.

2. Our Solution Hypersurface

Our solution is called Hypersurface and it consists of a dashboard, providing real-time insights for canteens, restaurants and other gastronomy businesses. The dashboard offers data visualization as well as predictions for the profit and revenue, users can choose from our available pre-trained models.

In order to implement our dashboard, we used the Django, which is a performant web framework written in Python.

Conclusion

As we worked our way through the different milestones of this project, to make a canteen business solution providing insights answering the client’s needs and bringing new solutions to their business.

A detailed business understanding was essential to come out with the right strategy to achieve each milestone. This process also required much of our the skills we learned this year to meet these objectives.