NOVA SCHOOL OF BUSINESS & ECONOMICS

2695 Introduction to Machine Learning
Project Description

Project Descriptions

- Project is a group based, semester-long assignment for students to solve a business question by performing machine learning on a medium-sized dataset.
- Each group consists of 3 students.
- Each group can choose up to 3 projects, ranked by the preferences (preferences are stated in the excel sheet) and we will do our best to accommodate everyone's interest.
- Project code will need to be submitted by the last lecture, and the video by Monday, May 12th
- Team members will have the option to evaluate each other's contribution.
- Attending the final class on Friday May 16th is mandatory for project discussion.

Project assignment timeline

1. A list of possible project ideas is given.
2. You can also choose your own topic and dataset, BUT the project proposal should be approved before proposal submission. This is done by sending an email to the TA & instructor with a brief project description. The approval must be obtained BEFORE Friday February 28th.
3. You should submit your top 3 project preferences on Moodle by Thursday, February 28th. The preference Excel sheet will be available on Moodle after class 2, on February 14th.
4. An assignment of students to projects will be available in week 5.
5. All students without a group will be assigned to groups based on their student number and project preferences. We will do our best to accommodate everyone’s interest, and we will solve ties by giving preference to earlier submissions.

Final Evaluation Criteria

1. What business problem are you solving
2. What machine learning problem are you solving
3. Explorative data analysis & data preparation
4. Machine learning models
5. Evaluating and interpreting machine learning models

Final evaluation will be based on:

- Submitted project notebook
- Video of the project presentation
- Question & Answer session where all the team members should be able to answer all project questions

Team member evaluation: lowering the grade of the non-contributing team member

- Each team member will have the possibility to evaluate other team members on their **overall participation and contribution to the project**. So, if they feel they all participated equally, they can grade each other with 20. This will be the default value assumed, if the team members do not grade each other.
- To get an individual project grade: project grade will be weighted with the average of the grade given by the other team members. Example:
  - The project gets a grade 18.
  - Two team members, A and B, give the third member C, the following grades: 19 (graded by student A) and 18 (graded by student B), as student C did not participate equally in the project. Hence the student C gets the following grade:
  - 18 (*project grade*) x ((18+19)/2)/20 (*member contribution*) =18 x 18.5/20= **16.65**
  - Note that if you grade a fellow teammate with 18 that does not mean that they should get an 18, only that whatever the final project grade is, it should be weighted with 18/20.
- **Hence, any grade less than 20 will lead to a decreased grade of a fellow team member**. Please be mindful of this when you evaluate each other!

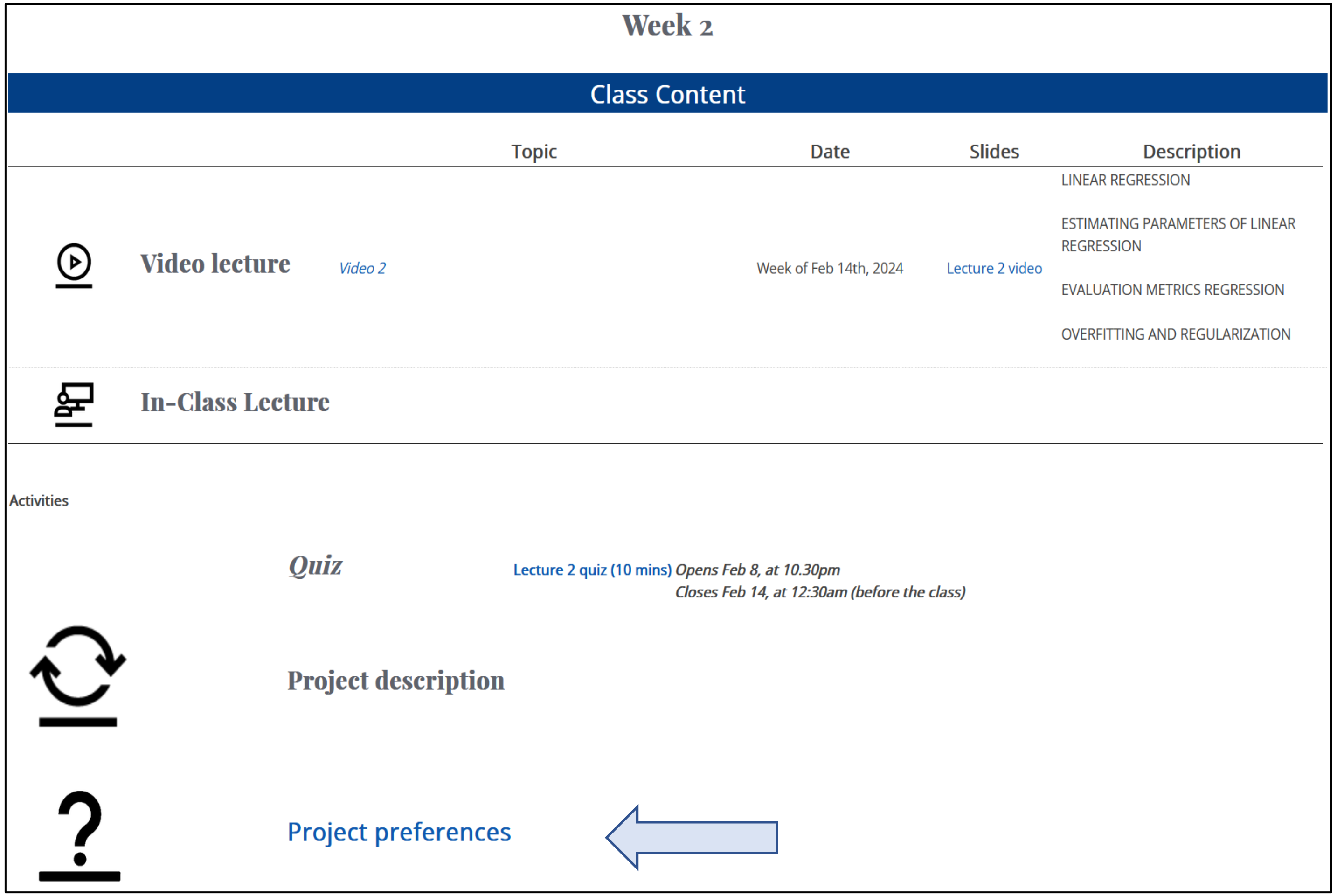
Team member evaluation: increasing the grade of the Most Valuable Team Member

- Each team member will have the possibility to nominate a team member if they feel that they have contributed more than the other team members.
- If a team member gets most of the team votes for being the MVTM, their grade may be increased.
- **Nominating MVTM is not mandatory, should be done for the cases when the workload was not equal.**

Project Preferences

STEP 1

- In week 2, click on Project preferences (link also available on the **main page**).



Week 2

Class Content

		Topic	Date	Slides	Description
Video lecture	*Video 2*		Week of Feb 14th, 2024	Lecture 2 video	LINEAR REGRESSION ESTIMATING PARAMETERS OF LINEAR REGRESSION EVALUATION METRICS REGRESSION OVERFITTING AND REGULARIZATION
In-Class Lecture					

Activities

Quiz Lecture 2 quiz (10 mins) *Opens Feb 8, at 10.30pm*
Closes Feb 14, at 12:30am (before the class)

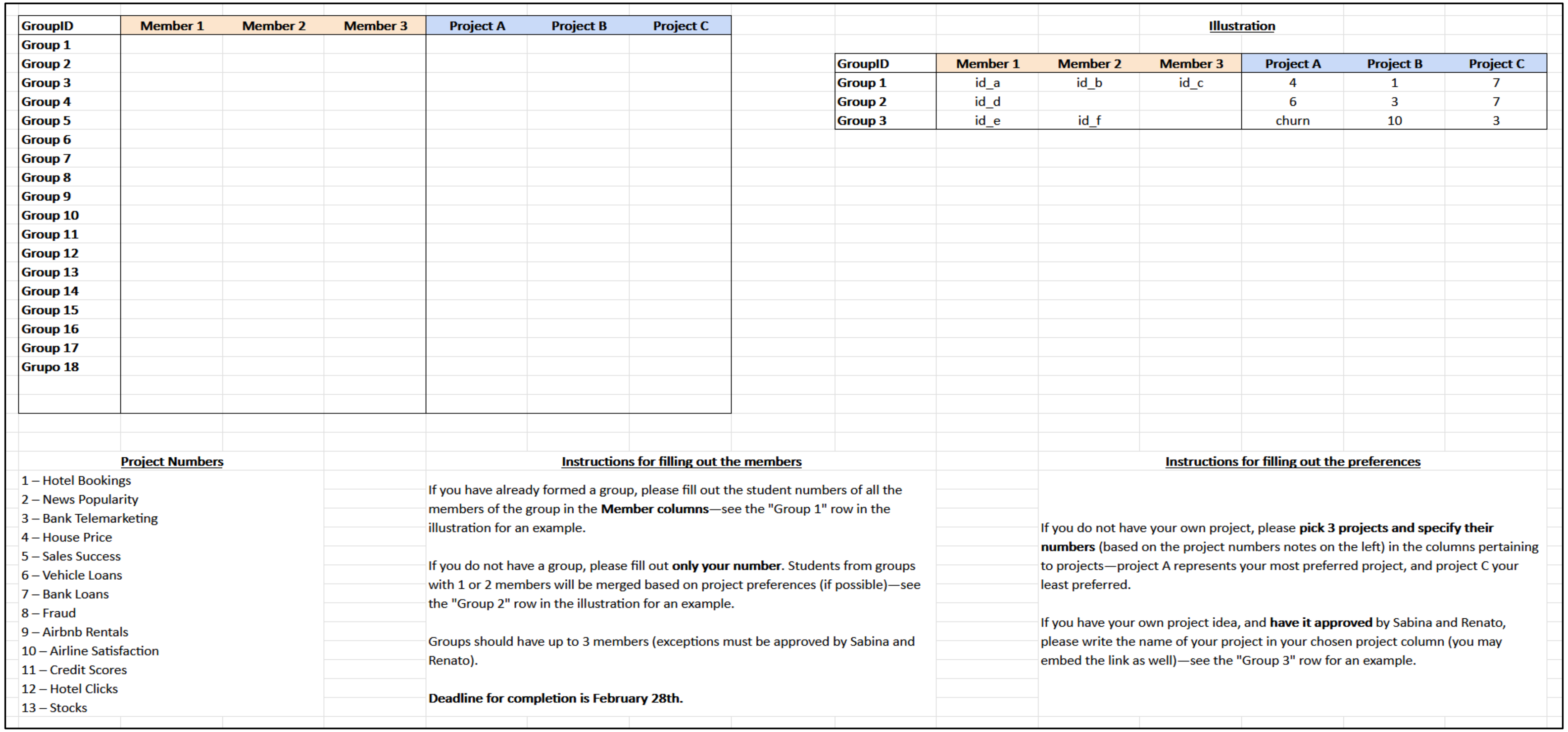
Project description

Project preferences

Project Preferences

STEP 2

- In the shared excel, follow the instructions below and write your project preferences.



GroupID	Member 1	Member 2	Member 3	Project A	Project B	Project C
Group 1						
Group 2						
Group 3						
Group 4						
Group 5						
Group 6						
Group 7						
Group 8						
Group 9						
Group 10						
Group 11						
Group 12						
Group 13						
Group 14						
Group 15						
Group 16						
Group 17						
Grupo 18						

Illustration

GroupID	Member 1	Member 2	Member 3	Project A	Project B	Project C
Group 1	id_a	id_b	id_c	4	1	7
Group 2	id_d			6	3	7
Group 3	id_e	id_f		churn	10	3

Project Numbers

1 – Hotel Bookings
2 – News Popularity
3 – Bank Telemarketing
4 – House Price
5 – Sales Success
6 – Vehicle Loans
7 – Bank Loans
8 – Fraud
9 – Airbnb Rentals
10 – Airline Satisfaction
11 – Credit Scores
12 – Hotel Clicks
13 – Stocks

Instructions for filling out the members

If you have already formed a group, please fill out the student numbers of all the members of the group in the **Member columns**—see the "Group 1" row in the illustration for an example.

If you do not have a group, please fill out **only your number**. Students from groups with 1 or 2 members will be merged based on project preferences (if possible)—see the "Group 2" row in the illustration for an example.

Groups should have up to 3 members (exceptions must be approved by Sabina and Renato).

Deadline for completion is February 28th.

Instructions for filling out the preferences

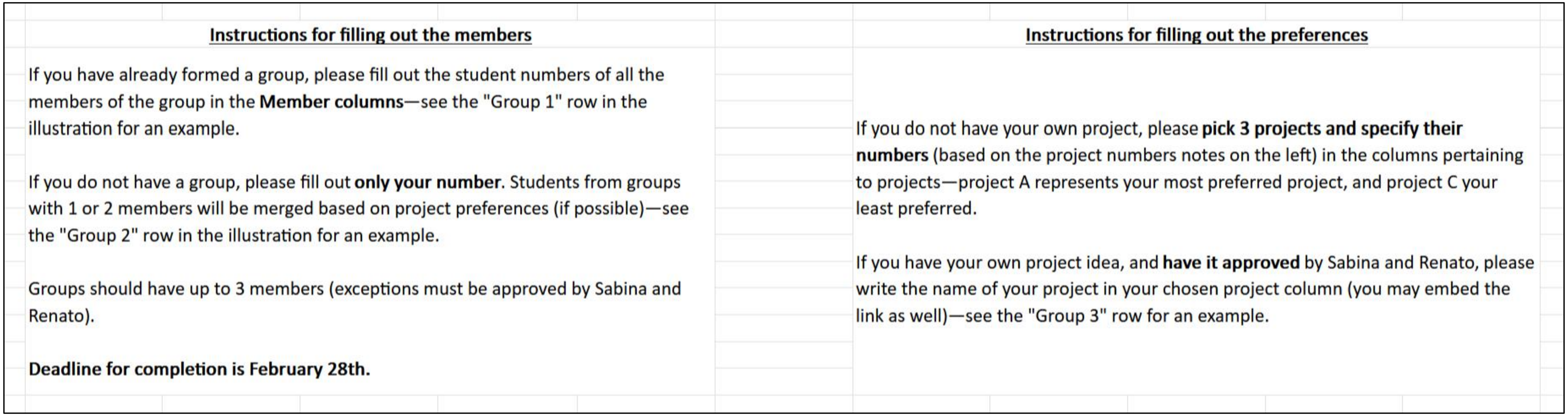
If you do not have your own project, please **pick 3 projects and specify their numbers** (based on the project numbers notes on the left) in the columns pertaining to projects—project A represents your most preferred project, and project C your least preferred.

If you have your own project idea, and **have it approved** by Sabina and Renato, please write the name of your project in your chosen project column (you may embed the link as well)—see the "Group 3" row for an example.

Project Preferences

STEP 2

- In the shared excel, follow the instructions below and write your project preferences.



Instructions for filling out the members

If you have already formed a group, please fill out the student numbers of all the members of the group in the **Member columns**—see the "Group 1" row in the illustration for an example.

If you do not have a group, please fill out **only your number**. Students from groups with 1 or 2 members will be merged based on project preferences (if possible)—see the "Group 2" row in the illustration for an example.

Groups should have up to 3 members (exceptions must be approved by Sabina and Renato).

Deadline for completion is February 28th.

Instructions for filling out the preferences

If you do not have your own project, please **pick 3 projects and specify their numbers** (based on the project numbers notes on the left) in the columns pertaining to projects—project A represents your most preferred project, and project C your least preferred.

If you have your own project idea, and **have it approved** by Sabina and Renato, please write the name of your project in your chosen project column (you may embed the link as well)—see the "Group 3" row for an example.

Project 1: Hotel bookings

- In tourism and travel related industries it’s important to forecast demand, understand seasonality trends and customer base.
- This data set contains booking information for a city hotel and a resort hotel, includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. It also has information on booking status, whether it was cancelled or not.
- Which booking will be cancelled?
- There are 100K instances and 32 variables.
- Dataset available at: Hotel Bookings

Project 2: Popularity of online news content

- Within the expansion of the Internet and Web 2.0, there has also been a growing interest in online news, which allow easy and fast spread of information around the globe. Predicting popularity of online news is valuable for authors, content providers, advertisers and even activists/politicians (e.g., to understand or influence public opinion).
- This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years.
- There are 40K instances and around 61 variables
- How popular will some content be? How many shares in social media will it get?
- Dataset available at: News Popularity

Project 3: Bank telemarketing

- To prioritize and select the next customers to be contacted during bank marketing campaigns, it is important to predict the outcome of the sales call. Contacting only those clients with higher probability of subscribing, leads to shorter time duration and lower campaign cost, as well as fewer and more effective phone calls, leading to lower client and employee stress.
- Dataset contains information related to direct marketing campaigns of a Portuguese retail bank, from May 2008 to June 2013. The marketing campaigns were based on phone calls. Each record includes the output target, the contact outcome and candidate input features, such as client information and social and economic context attributes.
- There are 45K instances and 19 variables.
- Which clients will subscribe a long-term deposit?
- Dataset available at: Bank Telemarketing

Project 4: House rental price

- The calculation of house prices is important in many areas of the market, including real estate, bank lending, and for tax purposes.

- This dataset contains house rental prices from Craigslist from all over the USA. Each record includes the details of the property such as its type (house, apartment,…) area, number of rooms, location, as well as the text description from the property listing.

- The dataset contains 100K instances and 22 variables.

- What will be the rental price of a property?
- Dataset available at: [House Prices]

Project 5: Successful sales campaigns

- Predicting successful sales opportunities enhances the understanding of the sales pipeline's flow and function. This insight allows businesses to optimize conversion strategies, forecast revenue more accurately, and pinpoint potential inefficiencies in their sales process.
- This dataset contains sales campaign data of an automotive parts wholesale supplier.
- The dataset contains 78K instances and 19 variables.
- Which sales campaign will result in a loss, and which will result in a win? What will be the value of the sales opportunity?
- Dataset available at: [Sales Success]

Project 6: Vehicle loans

- Financial institutions incur significant losses due to the default of vehicle loans. This has led to tightening up of vehicle loan underwriting and increased vehicle loan rejection rates.
- The dataset contains the information of the borrower (demographics such as age, Identity proof etc.), details of the loan and bureau data and history.

- The dataset contains 100K instances and 41 variables.
- Will the borrower default on a vehicle loan?
- Dataset available at: Vehicle Loans

Project 7: Bank loans

- Models for predicting whether an applicant should be accepted for a loan are widely used, and a very common implementation of machine learning practices.
- The goal is to predict which individuals will be accepted for a loan.
- 100K instances and 21 variables, some features are anonymized.
- Will an individual be accepted for a loan?
- Dataset available at: Bank Loans

Project 8: Fraud detection

- Another common implementation of machine learning in the credit space is detecting fraud. "It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase."

- 100K instances and 22 variables.

- Is this use of the credit card fraudulent?

- Dataset available at: Fraud

Project 9: Airbnb rentals

- Predicting Airbnb rental prices is crucial for optimizing revenue, ensuring competitive pricing, and meeting market demand efficiently.

- The dataset contains information such as listing name, latitude and longitude of listing, the neighborhood, price, room type, minimum number of nights, and city.

- 100K instances and 18 variables.

- What is the market price for this rental?

- Dataset available at: Airbnb Rentals

Project 10: Airline Passenger Satisfaction

- Predicting airline passenger satisfaction is important for enhancing customer loyalty and competitive differentiation in the airline industry. It enables carriers to proactively implement service improvements, tailor experiences, and efficiently allocate resources towards factors that increase overall traveler contentment and retention.

- The dataset contains information about each passenger, their flight, and type of travel, as well as their evaluation of different factors like cleanliness, comfort, service, and overall experience.

- 100K instances and 24 variables

- Will the airline passenger be happy with the service?

- Dataset available at: Airline Satisfaction

Project 11: Credit scores

- Predicting credit scores is important for financial institutions to assess risk accurately and to tailor credit offerings to individual risk profiles, optimizing capital allocation while expanding access to credit for qualified borrowers.

- The dataset contains information about bank clients, like their occupation and income, their ban portfolio and their credit score.

- 100K instances and 28 variables

- What is the credit score of each client?

- Dataset available at: [Credit Scores]

Project 12: Online hotel bookings

- Predicting online hotel booking outcomes is important for businesses to optimize occupancy rates and revenue by anticipating guest behavior and booking trends. It also allows for strategic pricing and targeted marketing efforts.

- The dataset contains information about the user, hotel and the search details such as number of people, check-in and check-out days.

- 100K instances and 24 variables

- Will the client book the hotel?

- Dataset available at: Hotel Clicks

Project 13: Stock market prediction

- Stock market prediction is crucial for investors and financial analysts to anticipate market movements, enabling informed investment decisions and risk management. It supports strategic portfolio allocation, maximizes returns, and minimizes losses by leveraging insights into future market trends and volatility.
- The dataset timeline of stock market information for various global brands.
- 100K instances and 12 variables
- Can we predict the stock price?
- Dataset available at: Stocks