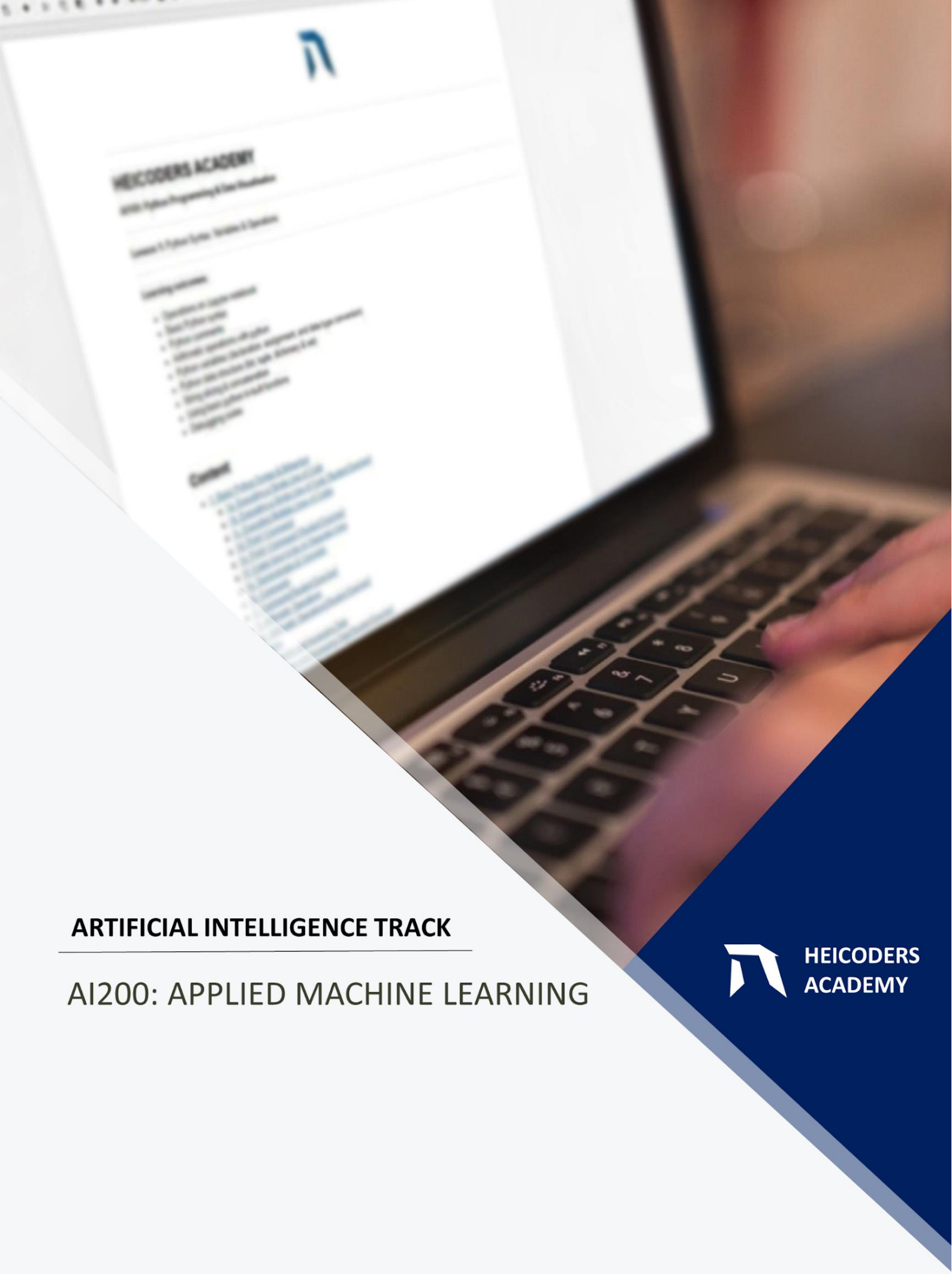


ARTIFICIAL INTELLIGENCE TRACK

AI200: APPLIED MACHINE LEARNING

HEICODERS ACADEMY

DEVELOPED BY: HEICODERS ACADEMY

ARTIFICIAL INTELLIGENCE TRACK

AI200: APPLIED MACHINE LEARNING
CAPSTONE PROJECT

1. Why Kaggle?

For the AI200 Capstone, students will work on a Machine Learning prediction project on the Kaggle platform. The primary reasons are two-fold:

1. Students are exposed to the end-to-end process of working on a Kaggle dataset for Machine Learning. This is crucial so that students are equipped to work on other Kaggle projects independently after AI200 to continue improving their skills and portfolio.

2. By the virtue of Kaggle being a widely known platform by data scientists worldwide, students will have a significant edge in career outcomes with the inclusion of Kaggle projects in their resume or portfolio that showcases their technical skills.

2. Business Scenario

You work for the LendingClub company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
    - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
    - Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are excluded from the dataset.
    - Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available within the company nor this dataset.

4. Business Objectives

LendingClub is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface. Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA and machine learning is the aim of this case study. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

5. Project Description

This In-Class Prediction Challenge is modelled after the LendingClub Issued Loans dataset. LendingClub is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. LendingClub is the world's largest peer-to-peer lending platform.

Solving this case study will give us an idea about how real business problems are solved using EDA and Machine Learning. In this case study, we will also develop an understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

In this competition, you'll be parsing through LendingClub's complete loan dataset and build a machine learning model to predict which of the loans are likely to be defaulted. Loan defaults is an expensive problem which any financial institute that engages in borrowing inadvertently faces. Each year, the financial industry loses billions of dollars due to loan defaults.

6. Datasets and Project Deliverables

You will require 2 data files for this competition:

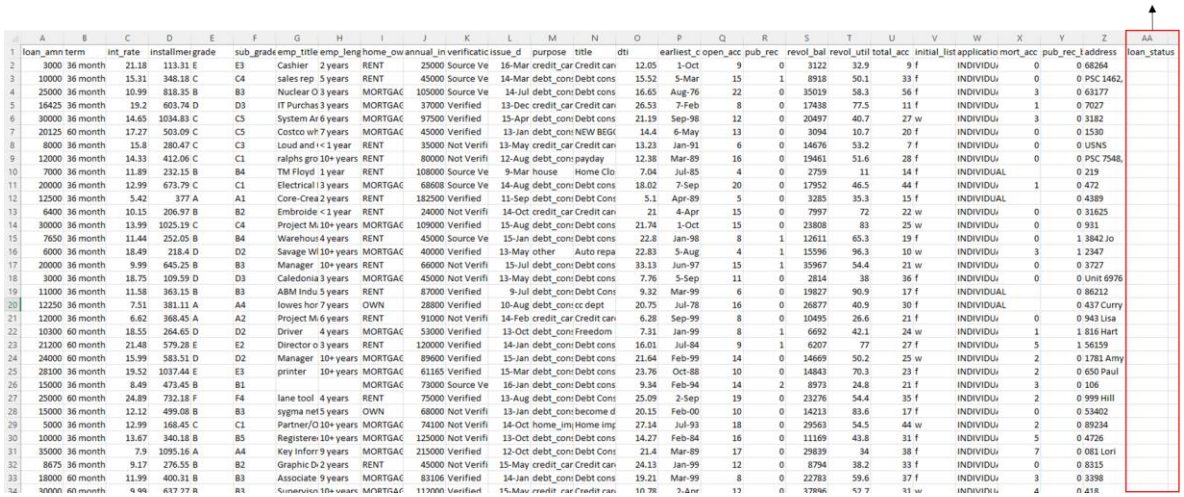
- lc_trainingset.csv contains the features and outcome column (**loan_status**) for loans made by LendingClub. You are to build a predictive classification model, using this training set to predict the outcome in **loan_status** column.
- lc_testset.csv. contains the same features but **has no outcome/labels** (i.e. **loan_status**)
    - You are to use the model trained from *lc_trainingset.csv* to predict loan_status from this test set, and submit your predictions in .csv format to Kaggle.
    - The ground truth (outcomes/labels) for this dataset is known only by the Kaggle platform. Kaggle will calculate the AUC score of your submitted predictions to determine your position on the leaderboard.

Deliverables (20 marks in total)

(1) Submit Jupyter Notebook that uses the provided datasets to
   a. Perform exploratory data analysis, data cleaning, feature engineering
   b. Build machine learning models and evaluate with AUC metric
   c. Generate prediction probabilities for the loan_status column

   (Total: 10 marks)

You are supposed to generate a prediction of the outcome for each row of the data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	loan_amn	term	int_rate	installmer	grade	sub_grad	emp_title	emp_leng	home_ow	annual_in	verificatic	issue_d	purpose	title	dti	earliest_c	open_acc	pub_rec	revol_bal	revol_util	total_acc	initial_list	applicatio	mort_acc	pub_rec_t	address	loan_status
2	3000	36 month	21.18	113.31	E	E3	Cashier	2 years	RENT	25000	Source Ve	16-Mar	credit_car	Credit can	12.05	1-Oct	9	0	3122	32.9	9	f	INDIVIDU	0	0	68264	
3	10000	36 month	15.31	348.18	C	C4	sales rep	5 years	RENT	45000	Source Ve	14-Mar	debt_con	Debt cons	15.52	5-Mar	15	1	8918	50.1	33	f	INDIVIDU	0	0	PSC 1462,	
4	25000	36 month	10.99	818.35	B	B3	Nuclear O	3 years	MORTGAC	105000	Source Ve	14-Jul	debt_con	Debt cons	16.65	Aug-76	22	0	35019	58.3	56	f	INDIVIDU	3	0	63177	
5	16425	36 month	19.2	603.74	D	D3	IT Purchas	3 years	MORTGAC	37000	Verified	13-Dec	credit_car	Credit can	26.53	7-Feb	8	0	17438	77.5	11	f	INDIVIDU	1	0	7027	
6	30000	36 month	14.65	1034.83	C	C5	System Ar	6 years	MORTGAC	97500	Verified	15-Apr	debt_con	Debt cons	21.19	Sep-98	12	0	20497	40.7	27	w	INDIVIDU	3	0	3182	
7	20125	60 month	17.27	503.09	C	C5	Costco wh	7 years	MORTGAC	45000	Verified	13-Jan	debt_con	NEW BEG(	14.4	6-May	13	0	3094	10.7	20	f	INDIVIDU	0	0	1530	
8	8000	36 month	15.8	280.47	C	C3	Loud and I	< 1 year	RENT	35000	Not Verifi	13-May	credit_car	Credit can	13.23	Jan-91	6	0	14676	53.2	7	f	INDIVIDU	0	0	USNS	
9	12000	36 month	14.33	412.06	C	C1	ralphs gro	10+ years	RENT	80000	Not Verifi	12-Aug	debt_con	payday	12.38	Mar-89	16	0	19461	51.6	28	f	INDIVIDU	0	0	PSC 7548,	
10	7000	36 month	11.89	232.15	B	B4	TM Floyd	1 year	RENT	108000	Source Ve	9-Mar	house	Home Clo	7.04	Jul-85	4	0	2759	11	14	f	INDIVIDUAL		0	219	
11	20000	36 month	12.99	673.79	C	C1	Electrical I	3 years	MORTGAC	68608	Source Ve	14-Aug	debt_con	Debt cons	18.02	7-Sep	20	0	17952	46.5	44	f	INDIVIDU	1	0	472	
12	12500	36 month	5.42	377	A	A1	Core-Crea	2 years	RENT	182500	Verified	11-Sep	debt_con	Debt Cons	5.1	Apr-89	5	0	3285	35.3	15	f	INDIVIDUAL		0	4389	
13	6400	36 month	10.15	206.97	B	B2	Embroide	< 1 year	RENT	24000	Not Verifi	14-Oct	credit_car	Credit can	21	4-Apr	15	0	7997	72	22	w	INDIVIDU	0	0	31625	
14	30000	36 month	13.99	1025.19	C	C4	Project M	10+ years	MORTGAC	109000	Verified	15-Aug	debt_con	Debt cons	21.74	1-Oct	15	0	23808	83	25	w	INDIVIDU	0	0	931	
15	7650	36 month	11.44	252.05	B	B4	Warehous	4 years	RENT	45000	Source Ve	15-Jan	debt_con	Debt cons	22.8	Jan-98	8	1	12611	65.3	19	f	INDIVIDU	0	1	3842 Jo	
16	6000	36 month	18.49	218.4	D	D2	Savage Wi	10+ years	MORTGAC	40000	Verified	13-May	other	Auto repa	22.83	5-Aug	4	1	15596	96.3	10	w	INDIVIDU	3	1	2347	
17	20000	36 month	9.99	645.25	B	B3	Manager	10+ years	RENT	66000	Not Verifi	15-Jul	debt_con	Debt cons	33.13	Jun-97	15	1	35967	54.4	21	w	INDIVIDU	0	0	3727	
18	3000	36 month	18.75	109.59	D	D3	Caledonia	3 years	MORTGAC	45000	Not Verifi	13-May	debt_con	Debt cons	7.76	5-Sep	11	0	2814	38	36	f	INDIVIDU	0	0	Unit 6976	
19	11000	36 month	11.58	363.15	B	B3	ABM Indu	5 years	RENT	87000	Verified	9-Jul	debt_con	Debt Cons	9.32	Mar-99	6	0	19827	90.9	17	f	INDIVIDUAL		0	86212	
20	12250	36 month	7.51	381.11	A	A4	lowes hor	7 years	OWN	28800	Verified	10-Aug	debt_con	cc dept	20.75	Jul-78	16	0	26877	40.9	30	f	INDIVIDUAL		0	437 Curry	
21	12000	36 month	6.62	368.45	A	A2	Project M	6 years	RENT	91000	Not Verifi	14-Feb	credit_car	Credit can	6.28	Sep-99	8	0	10495	26.6	21	f	INDIVIDU	0	0	943 Lisa	
22	10300	60 month	18.55	264.65	D	D2	Driver	4 years	MORTGAC	53000	Verified	13-Oct	debt_con	Freedom	7.31	Jan-99	8	1	6692	42.1	24	w	INDIVIDU	1	1	816 Hart	
23	21200	60 month	21.48	579.28	E	E2	Director o	3 years	RENT	120000	Verified	14-Jan	debt_con	Debt cons	16.01	Jul-84	9	1	6207	77	27	f	INDIVIDU	5	1	56159	
24	24000	60 month	15.99	583.51	D	D2	Manager	10+ years	MORTGAC	89600	Verified	15-Jan	debt_con	Debt cons	21.64	Feb-99	14	0	14669	50.2	25	w	INDIVIDU	2	0	1781 Amy	
25	28100	36 month	19.52	1037.44	E	E3	printer	10+ years	MORTGAC	61165	Verified	15-Mar	debt_con	Debt cons	23.76	Oct-88	10	0	14843	70.3	23	f	INDIVIDU	2	0	650 Paul	
26	15000	36 month	8.49	473.45	B	B1			MORTGAC	73000	Source Ve	16-Jan	debt_con	Debt cons	9.34	Feb-94	14	2	8973	24.8	21	f	INDIVIDU	3	0	106	
27	25000	60 month	24.89	732.18	F	F4	lane tool	4 years	RENT	75000	Verified	13-Aug	debt_con	Debt Cons	25.09	2-Sep	19	0	23276	54.4	35	f	INDIVIDU	2	0	999 Hill	
28	15000	36 month	12.12	499.08	B	B3	sygma net	5 years	OWN	68000	Not Verifi	13-Jan	debt_con	become d	20.15	Feb-00	10	0	14213	83.6	17	f	INDIVIDU	0	0	53402	
29	5000	36 month	12.99	168.45	C	C1	Partner/O	10+ years	MORTGAC	74100	Not Verifi	14-Oct	home_im	Home imp	27.14	Jul-93	18	0	29563	54.5	44	w	INDIVIDU	2	0	89234	
30	10000	36 month	13.67	340.18	B	B5	Registered	10+ years	MORTGAC	125000	Not Verifi	13-Oct	debt_con	Debt cons	14.27	Feb-84	16	0	11169	43.8	31	f	INDIVIDU	5	0	4726	
31	35000	36 month	7.9	1095.16	A	A4	Key Infor	9 years	MORTGAC	215000	Verified	12-Oct	debt_con	Debt Cons	21.4	Mar-89	17	0	29839	34	38	f	INDIVIDU	7	0	081 Lori	
32	8675	36 month	9.17	276.55	B	B2	Graphic D	2 years	RENT	45000	Not Verifi	15-May	credit_car	Credit can	24.13	Jan-99	12	0	8794	38.2	33	f	INDIVIDU	0	0	8315	
33	18000	60 month	11.99	400.31	B	B3	Associate	9 years	MORTGAC	83106	Verified	14-Jan	debt_con	Debt cons	19.21	Mar-99	8	0	22783	59.6	37	f	INDIVIDU	3	0	3398	
34	30000	60 month	9.99	637.27	B	B3	Supervisa	10+ years	MORTGAC	112000	Verified	15-May	credit_car	Credit can	10.78	2-Apr	12	0	37896	52.7	31	w	INDIVIDU	4	0	418	

(2) Submit predictions from your final model in .csv format to the Kaggle competition
(Total: 5 marks)

Marks	AUC Score
5 marks	0.89+ (Exceptional)
4 marks	0.7+ (Good)
3 marks	0.6+ (Satisfactory)
2 marks	>0.5 (Fair)
1 mark	0.5 or below (Submission)

(3) Additionally, all students are to individually submit reflections on the prompts below. (Total: 5 marks) *Submission form will be available on eLearn after Lesson 7.*

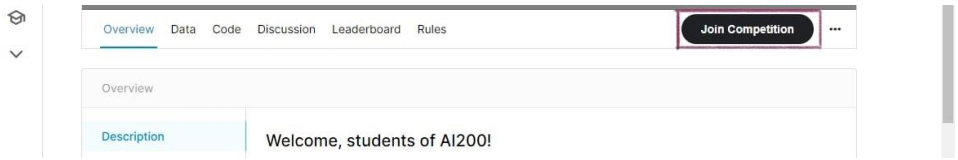
- Your key takeaways and insights from working on this project (2 mark)
- What are some areas in your current workplace where you think the introduction of big data application will be beneficial? (2 mark)
- Walk us through how you will go about implementing one big data application in your workplace (1 mark)

7. Pre-Project Task: Make Your First Kaggle Submission (Due on Lesson 6)

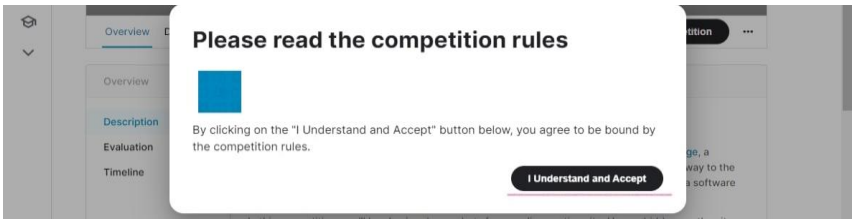
Please complete the following steps before Lesson 6:

1. Go to the Kaggle Competition website provided on eLearn.

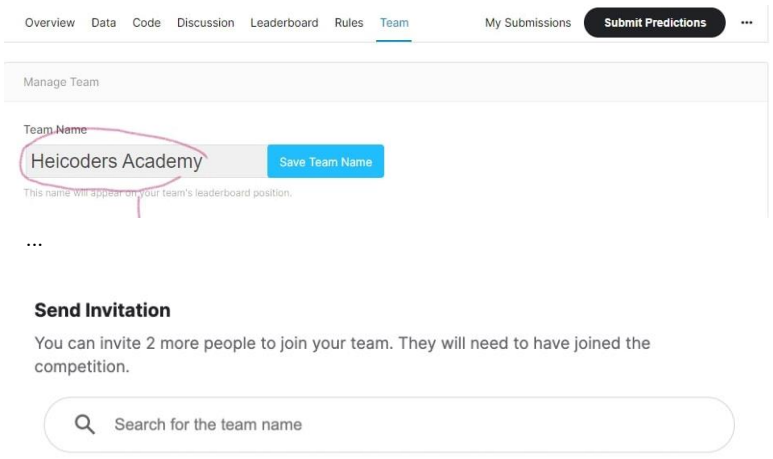
2. Click 'Join Competition' (see screenshot)



3. Read the competition rules and click "I Understand and Accept'



4. After all your teammates have joined the competition, go to the "Team" tab. One representative will collect the "Team Name" of his/her teammate(s), and send them the invitation to merge team. Make sure all teammates approve your request before moving on to the next step.



5. Once all team members are added & your team is complete, change your team name to reflect your Team Number to facilitate grading (e.g. Team X).

6. As a team, make your first Kaggle submission. (Only 1 team member needs to do this)

    a. Click Submit Predictions and upload the sample-submission.csv from eLearn.
    b. The description field is only visible to the instructor team and your teammates – this field is for you to keep track of which submission is for which model
        i. Example: *"xgboost with 20 features and max_depth=__"*
        ii. For this sample submission, you can simply fill in *"sample submission"*.

7. Now, your team name/number should appear on the Public Leaderboard. Do PM your teaching assistants if you encounter any issues or need further assistance!

8. Competition Rules

Submission Limit	Each team may submit a maximum of 6 entries per day. You may select up to 2 final submissions for judging. The better of the two will be counted towards your final AUC score.
Eligibility	The Competition is open to all AI200 students registered in the current cohort. Submissions must only be made by the same teams as provided by the AI200 instructor team on eLearn.
Winners' Obligations	As a condition of receipt of the prizes, winners must: • Upload the code behind the final model to Heicoders eLearn in the form of a Jupyter Notebook before Lesson 8, which should allow the instructor team to reproduce the winning submission for verification. • Present their winning submission Notebook to the class on Lesson 8 (estimated duration of 5 minutes)
No Sharing of Codes / Data	Sharing code or data outside of teams is not permitted.
One Account per Participant	As Kaggle strictly prohibits signing up from multiple accounts, no participant may submit from multiple accounts. If discovered by Kaggle, this may lead to permanent deactivation and suspension of affected Kaggle accounts.
Determining Winners	This Competition is a challenge of skill, and the results are determined solely by leaderboard ranking on the private leaderboard at the end of the competition (subject to compliance with Competition Rules). Participants' scores and ranks on the public leaderboard are based on the AUC metric and determined by applying the predictions in the Submission to the ground truth of a 30% subset of the hidden test set outcomes used to generate the private leaderboard. Prize awards are subject to verification of eligibility and compliance with these Competition Rules. All decisions of the Competition Sponsor and judges will be final and binding on all matters relating to this Competition. Competition Sponsor reserves the right to examine the Submission and any associated code or documentation for compliance with these Competition Rules. If the Submission demonstrates a breach of these Competition Rules, Competition Sponsor may disqualify the Submission(s) at its discretion.
Resolving Ties	A tie between two or more valid and identically ranked submissions will be resolved in favour of the tied submission submitted first.
Declining Prizes	A Participant may decline to be nominated as a Winner by notifying Heicoders directly within 1 day following the Competition deadline, in which case the declining Participant forgoes any prize or other features associated with winning the Competition. Kaggle reserves the right to disqualify a Participant who so declines at Kaggle's sole discretion if Kaggle deems disqualification appropriate.