**Task Description: IronHacks Data Science Training Challenge Fall 2022**

Missed a Live Session? [Go to Youtube for the recap!](https://www.youtube.com/watch?v=QaNo7hQzOl8&list=PLIT-Ltqt_HO7xkuZD8svqPjl_TSIq-DKF)

**1. Context and Goal**

Using training data to predict unemployment claims in the State of Indiana!

We ask you to predict the number of unemployment claims using historical data provided by DWD, as well as demographics data. You are given variables such as race, education, categories of employers, week number, etc. Your task is to predict the total unemployment claims for a specific census tract in Indiana in a specific week. Using such data, your predictions will inform citizens as well as decision-makers like the State of Indiana, the Department of Workforce Development, and others about the potential impact of unemployment to the state.

Please review the [rules](https://ironhacks.com/hacks/fall-2022/rules) before you start hacking.

**2. Data Science Task: Prediction modeling of unemployment claims in Indiana**

Your major task is to build a statistical model using data collected since week 1 of 2022 to predict the unemployment claims for census tracts in Indiana for future weeks. In other words, you have to predict the total\_claimsfor a future week. (Note, you may not need all those historical data in your modeling). This challenge works with real-time updates, so as you are moving through the challenge, you will have to make predictions for later weeks. You start with **week 39**, and you end with **week 43**.

In our [tutorials](https://ironhacks.com/hacks/fall-2022/tutorials), you will learn how to create a data model, but for your submissions, you can use a statistical model of your choice to make those predictions using historical data that we will provide you with. In fact, we ask you to **explore** a variety of models - from different kinds of regressions, time-series models, deep-learning models, and support vector machines. There is no limit to your effort to experiment. Indeed, we want to see a VARIETY of models being used.

Note: Even if you are not an expert in modeling temporal data at the start of the training you should not shy away from participation. That's the whole point of the training! A visual inspection of the data and some exploratory data analysis will indicate that even a simple model may offer a good starting point. To make it up to the highest rank of the participants, we encourage you to explore and try different models! This is a fantastic learning opportunity.

**3. Data and Libraries**

**3.1. Real-time data and data updates**

To make those predictions, we are providing pre-processed data from our data providers, the Department of Workforce Development (DWD) so that you can focus on the actual work of modeling. The level of analysis, and the most granular resolution of the data is a census tract. In our data, they can be identified with a unique tractnumber.

There are two tables that you can work with to train your model, called unemployement dataand wage data. These two tables provide you with the list of uu\_idfor which you should make the prediction. The schema of the pre-processed data that you will be used to "train" your model (e.g. estimating the coefficient for your regression). There is a **third** table, called prediction\_list, which contains the list of uu\_id, that is, tractsfor which you should make a prediction of unemployment for a particular week. This table has two columns: uu\_id, and week\_numberPlease note that the week\_number is changing over the course of this data science challenge. Our data is updated **real-time**, and so every week the prediction week changes.

| **Submission Nr** | **Training weeks** | **Prediction week** |
| --- | --- | --- |
| 1 to 3 | week 1 to week 37 2022 | 39 |
| 4 to 6 | week 1 to week 37 2022 | 40 |
| 7 to 9 | week 1 to week 37 2022 | 41 |
| 10 to 11 | week 1 to week 37 2022 | 42 |
| 12 to 14 | week 1 to week 37 2022 | 43 |

Do not forget that every week in our challenge for every third submission, you will work with a different dataset!

The schema of these tables is discussed in this [sheet](https://docs.google.com/spreadsheets/d/e/2PACX-1vTw11QCzmLmg1RHtwd3Xs4FpFP3zzZEr7-BKJ-hceqxYInKVsMyBkD20BySWDA1sr2p4yhpwsyY8SLl/pubhtml).

| **Nr** | **Table name** | **Brief Description** |
| --- | --- | --- |
| 1 | unemployment\_data | This table is the **primary** table for your modeling efforts. This table has 25 columns and includes the historical data from week 1 to the most recent week for the total\_claimsfor census tracts in Indiana. It also contains additional metrics such as gender, race, education level, and top industries with respect to claims |
| 2 | wage\_data | This table contains average wage information by census tract. |
| 3 | prediction\_list | This list provides the list of uu\_ids needed to make a prediction for the particular week week\_numberThis table does not contain any additional data. |

**3.2. Data access**

These tables are stored in BigQuery in a project called: ironhacks-data. In this project, you will find the dataset ironhacks\_competition. In there you find the tables with the names listed above.

[Tutorial 2](hptts://ironhacks.com/hacks/fall-2022/tutorials) describe how you can query data from Google Big Query. So if you have not looked into this tutorial, it is time to do this immediately! You should run the notebook in the tutorial, and also watch the video that has been posted. If you have questions, you can get in touch with us via the "Get in Touch" button on the top right hand corner. We are happy to set-up a one-to-one session if needed.

**3.3 Libraries**

We have pre-installed most of the eligible [libraries in python](https://docs.google.com/spreadsheets/d/1iIbhE5Px3mb2qdqRlnsdEUtnRL90oxm73YqMoSCXeXw/edit?usp=sharing) that you can use throughout the competition to work with the data in the JupyterLab.

Only the libraries Tensorflow, Keras, and Pytorch are not pre-installed. Please check the [Submission Template](https://ironhacks.com/notebook-viewer?path=https://raw.githubusercontent.com/ironhacks/Tutorials-COVID-19/master/tutorials-fall-2022/python/ironhacks_submission_template.ipynb) and cell two. **With the command !python3 -m pip install [package name]you can install these large libraries into your own "container". Once you install the libraries you'd like, restart your kernel in your Workspace and you will be all good to go!** As stated in the [rules section](https://ironhacks.com/hacks/fall-2022/rules) you are asked to use those libraries.

**4. Submission Requirements**

You will submit your results via the [submission page](https://ironhacks.com/hacks/fall-2022/submissions). Please carefully read the submission form when submitting to make sure you have everything correct.

**4.1. Submission Deadlines and Mandatory Submissions**

In this hack, there are **nine** submission opportunities, and **three mandatory** submission times at which you have to make a submission. See the table below for more details.

| **Submission Nr** | **Date** | **Day** | **Target prediction week** |
| --- | --- | --- | --- |
| Submission Opportunity 1 | 11/03 | Thursday | week 39 |
| Submission Opportunity 2 | 11/05 | Saturday | week 39 |
| Submission Milestone3 | 11/07 | Monday | week 39 |
| Submission Opportunity 4 | 11/10 | Thursday | week 40 |
| Submission Opportunity 5 | 11/12 | Saturday | week 40 |
| Submission Milestone 6 | 11/14 | Monday | week 40 |
| Submission Opportunity 7 | 11/17 | Thursday | week 41 |
| Submission Opportunity 8 | 11/19 | Saturday | week 41 |
| Submission Milestone 9 | 11/21 | Monday | week 41 |
| Submission Opportunity 10 | 11/23 | Wednesday | week 43 |
| Submission Milestone 11 | 11/28 | Monday | week 44 |
| Submission Opportunity 12 | 12/1 | Thursday | week 45 |
| Submission Opportunity 13 | 12/3 | Saturday | week 46 |
| Submission Milestone 14 | 12/5 | Monday | week 47 |

**4.2. Required Submission Material for Submissions**

For your submissions, you are asked to submit two files, a .csvand .ipynbfile, and complementary information (such as a model description and a selection of tags)

**4.2.1. File one: The .csvfile with your Predictions**

Your .csvfile with the prediction for the census tracts in Indiana with a unique uu\_idfor a specific  week\_numberis the most important part of your submission. It needs to follow the following structure: A comma-separated CSV file containing all the uu\_ids listed in the tableprediction\_list. Please include the headers in your .csvfile as follows.

uu\_id,week\_number,total\_claims

9bee42b55f891413ae7fac2d9d89280a,39,3

...

You need to save this file with the following filename: submission\_prediction\_output.csv.

**4.2.2 File two: A Jupyter Notebook (.ipynb) to Produce Predictions**

In addition to the .csvfile, you are also required to submit a Jupyter Notebook (.ipynb) that you used to produce the results in the .csvfile. It should contain the code that you have used to produce the predictions. We recommend that you also comment on your code. Please save it with the following name submission\_prediction\_output.ipynb.

**4.2.3. Complementary Information: Model Summaries, Tags, and References**

When you submit you are also asked to provide additional comments such as a description of your model (min 500 words), a list of references(working website links and doi is required) as well as a list of tags. In addition, you are asked to give credits to your peers in your group. If your new submission is inspired by the work of others, you should pick their pseudonamefrom a dropdown list. Please review the submission details carefully about how to write a proper model description.

**4.3. Submission Preparation**

To help you prepare a high-quality submission that meets technical (e.g. formatting of .csvand .ipynb) as well as non-technical requirements (e.g. submission summary), we have prepared two [tutorials](https://ironhacks.com/hacks/fall-2022/tutorials) for you: Tutorial 6a and tutorial 6b. Please check it out so that your submission is successfully considered for evaluation.

In this hack, you are required to use the IronHacks workspace (the workspace button is located on the upper right hand corner) to prepare the submission (see rules) so that all participants work with the same conditions, packages, and settings. We want this to be a fair competition. So if you are not running your code in our hub, you may lose points as your effort in the hub will count towards your final performance. Please make sure that you use our template notebook as you are working in our workspace. You can find the template [here](https://ironhacks.com/notebook-viewer?path=https://raw.githubusercontent.com/ironhacks/Tutorials-COVID-19/master/tutorials-fall-2022/python/ironhacks_submission_template.ipynb).

**IMPORTANT**: The detailed submission requirements are spelled out on the submission page. So, make sure you read carefully!

**5. Evaluation**

You will be evaluated on two **EQUALLY** weighted performance dimensions. We will discuss them next.

**5.1. Evaluation Category I: Prediction Accuracy**

The first evaluation metrics relate to the accuracy of your prediction. Each submission will be evaluated with the [**Mean Squared Prediction Error (MSPE)**](https://en.wikipedia.org/wiki/Mean_squared_error) as well as the [**Mean Absolute Prediction Error (MAPE)**](https://en.wikipedia.org/wiki/Mean_absolute_error) of your predicted values for total\_claimsfor each of uu\_ids listed in the prediction\_listfor the a. The formula of the **MSPE** and the **MAPE** are as follows:

MSPE=1n∑i=1n(yi−y^i)2MSPE=n1​i=1∑n​(yi​−y^​i​)2MAPE=1n∑i=1n∣yi−y^i∣MAPE=n1​i=1∑n​∣yi​−y^​i​∣

nn is the number uu*ids in the prediction\_list*yy*is the observed total\_claimsfor the prediction week of that particular submission. ~hat y*i~ is the predicted total\_claimsfor that prediction week of that particular submission.

We calculate your prediction accuracy right after your submission. The results should be visible very quickly - we plan to have your scores up within a day or so! Please be patient - if the scores are not published in the navigation section under the results tab, do not worry! The team might be reviewing your results. Keep checking on your scores!

A note: if you have log-transformed your data before estimating your coefficient, do not forget to revert that transformation before you submit your results!

**5.2. Evaluation Category II: Exploration**

There is a secondary and equally important evaluation metric that we call "exploration". Yes, this training is not just about building the best model. We want you to explore new models that you have not yet tried before. So while we want you to build more predictive models, we also want to make sure we recognize your effort to explore different models.

* Your effort to improve your statistical model (e.g. the number of value-adding submissions, the amount of value-adding changes to your software code without changing your model, number of unique commands)
* Your effort to try different solution approaches (e.g. your attempt to try out different functions and models in the statistical packages we have provided you with, or the coding of your own functions and models)
* Your activity in the workspace (e.g. how actively you work on the hub using the JupyterLab and BigQuery)

Such experimentation and exploration will not be evaluated in real-time as it takes some data pre-processing and expert judgment that cannot be performed in real-time. It will only be examined at the end of the contest. However, we will release a proxy measure for that after each of your submissions.

So isn't this exciting? You get the opportunity to explore different models?

**6. Prizes and incentives**

We will be evaluating individuals based on the two metrics stated above. Please review the scoring section for more details about how we rank individuals. At a minimum, prizes for individuals will be allocated as follows:

**6.1. Prizes (monetary)**

Top-scoring participants who submit in accordance with the Rules will be eligible for a share of **$2000**. In this hack, you will be awarded based on your individual performance. We evaluate each individual on their individual performance based on the **average rank** in the two categories mentioned above (accuracy and exploration). Please remember that for exploration we consider all submissions you made so the more submissions you made the more likely that you will score high. We will rank order the participants in each category and then calculate the **average rank** based on both ranks: prediction rank and exploration rank.

Please note that we split the overall hack into groups. Why? If the contest is smaller, the chances of winning increase. The goal is to have a size of about maximum 50 participants in each group. So that increases your chances of winning a lot (if you compare it to a Kaggle-style competition where you have much more competitors). In the past, we had about 30 people in each group.

Based on the rank, the cash prizes for individuals will be allocated as follows:

1. The **First Place** will receive USD 400
2. The **Second Place** will receive USD 200
3. The **Third Place** will receive USD 100
4. The **Forth Place** will receive USD 60
5. The **Fifth Place** will receive USD 40

In addition, there are more additional benefits for you:

**Every** participant who submits 9 times and fills in all surveys receives $10. So even if you are not among the best but you are just trying to use this challenge to learn, you earn money.

**6.2 Additional incentives that make you stand out**

In addition, there are more additional benefits for you:

* Certificate: Every participant will receive a digital certificate for participation with the logos of the sponsors.
* Showcases: We will publish selected models [www.ironhacks.com](http://www.ironhacks.com) at the end of the competition. See our current [showcases](https://ironhacks.com/hacks/showcase).
* Learning experience: Based on what we have learned from the past, IronHacks will offer a great learning experience and practice for data scientists.
* Job readiness: You can use the project you have build in future job applications. We are happy to work with you on this.