Myth-Busting the Mystical Service Market

How much of your data is contained in the (Co-)stars?

Aditi Chegu, Olivia Fang, Tejan Patel

Mentored by: Dan Shumow, Junaid Hasan

Astrological Apps

Astrology is a divination system which attempts to analyze an individual's personality and make predictions based on the location of celestial objects in the sky at the place and time of an individual's birth. Astrology classifies people based on their natal charts; this is the principle type of data we are analyzing. We are looking at astrological data, as it would be held by astrology apps, such as '*Co-Star*,' or '*The Pattern*'. The goal is to analyze whether or not it is possible that astrology apps are training machine learning models on advertiser data, and if this could leak Personally Identifiable Information.

Natal Charts

A natal chart (or birth chart) contains information about the location of planets in the sky at the time someone is born. The sky is divided into 12 regions that are assigned a house and sign. Each planet (including moon and sun) will be found in one of these regions and will then be assigned the region's house and sign. Given the location of each of the planets, we can construct a specific natal chart.

What is Co-Star?

In recent years, Co-Star, The Pattern, and other astrological apps have risen in popularity. Since its introduction in 2017, Co-Star has had over *20 million* downloads.

Astrology apps promise a variety of services such as daily predictions, compatibility charts, and interactions with astrologers.

These apps have raised *concerns about privacy* as they require Personally Identifiable Information such as a person's date of birth, location of birth, current location, and email address.

Our Hypothesis

We believe that Co-Star, under the guise of combining astrology with machine learning, gives its accurate predictions solely by exploiting its users' Personally Identifiable Information.

First, to verify that Co-Star is exploiting user information, we conducted an experiment to show that it is only possible for machine learning models to recall information accurately when overtrained on user data.

Then, we checked if it is possible to make similarly accurate predictions for users who are not part of the database of registered users.

To do this, we experimented with a dataset containing individual birthdays, employment duration, income, property ownership, household size, etc.

Last, we wanted to see if it was possible to predict– to a reasonable degree of accuracy– the information of a user given a partially filled natal chart. In doing this, we would be able to tell how much information was required by the app to identify a user.

A Classification Task: Car Ownership

Most users on Co-Star are a part of the training data of the ML model that makes predictions. Therefore, the predictions are fairly accurate. However, we wanted to see if the model is able to make similar accurate predictions for non-users. To do this, we inspected an overtrained model to see its performance on test data with and without astrological data.

Methods Used

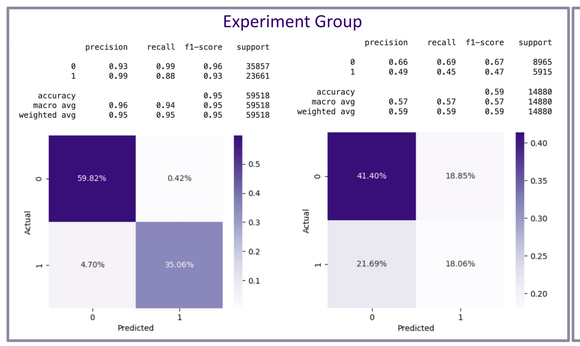
- Decision tree
  - builds a tree based on certain rules on the features.
- Random forest
  - uses multiple decision trees to build a more robust classifier.

Model Training Overview:

- We build a control model "Control group" with non-astrological features that are correlated with car ownership.
- Our "Experiment group" is a model that has all the features in the control model in addition to having astrology-related information.

Overfitting:

- The Experiment group experienced significant *overfitting* with a high training accuracy of **95%** but a much lower test accuracy of **59%.**
- The Control group, conversely, maintained a consistent accuracy of **69%** across both training and test datasets, showing no signs of overfitting.

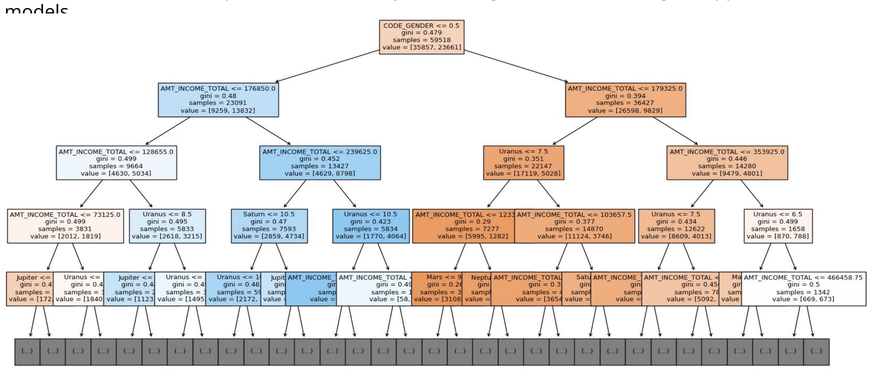


Hyperparameter Tuning:

- Conducted grid searches to optimize hyperparameters and reduce overfitting, particularly for the Control group using cross validation.
- With hyperparameter tuning, both the Control and Experiment models eventually converged to the same test accuracy of 69%.

Analysis and Interpretation:

- The discrepancy in overfitting between the Control and Experiment groups is noteworthy, especially since it does not seem to be related to the number of features (dimensionality).
- The consistent performance of the Experiment group suggests that the inclusion of astrology-related information might be influencing the model's generalization ability with training data, but not predicting ability with test data.
- If an astrological application predicts specific user features with more than 68% accuracy, it likely indicates overfitting, possibly due to using data other than just birthdays.
- Achieving around 68% accuracy, which matches the highest model accuracy observed in this study, can be achieved with or without astrological data, implying that the application might be utilizing additional types of data beyond birthdays.
- These behaviors underscore the importance of critically evaluating what data astrological apps use and how they build their predictive models.



A Regression Task: Invert Birth Chart

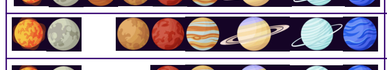
Most often, people use their birthdate to get their Natal Chart. The natal chart is then used to make predictions and draw insights into their life and personality. In this machine learning experiment, we analyze *if the inverse is possible*. *Can we predict a birth date given a natal chart?*

We conducted this experiment on a uniform sample of birth days and birth charts with a cardinality of 700,000 over the last 200 years.

In our first experiment, we used *scikit-learn's decision tree regression* to predict age in days, given a star chart. Recall, planetary signs are derived from star positions at a person's time of birth. This model performed very well with a Mean Squared Error (MSE) of 1.7 days on our testing set. This shows that indeed it is possible to predict the age of a person fairly accurately given their star chart.

We asked a deeper question, can one predict the date of birth with a partial start chart? In other words, given the location of a subset of planets, is it possible to guess a person's date of birth?

For this, we performed an iterative run, where at every iteration, we removed one planet at a time to see which planet could be discarded from the model, while preserving the best prediction possible. In the end, we have found the Sun, Mars, Jupiter, Saturn. and Neptune as the most efficient and effective at predicting birth date. Using these planets, we achieved an MSE of 48 days or around a month and a half. It can be concluded that these five celestial bodies provide the most valuable information in determining one's birthdate. Below, you can see our sequential elimination of planets.

Planets Used	MSE (days)
	1.7
	3.3
	16.0
	17.6
	48.5

Conclusion & Future Work

From the experiments, the evidence suggests that Co-Star uses Personally Identifiable Information to give accurate predictions to users. In the first experiment, we showed astrology is only viable to predict personal information when the ML model overfits the training data.

We also concluded that the information of five planets is enough to reasonably identify (48 days) a person's birthdate and ultimately a user.

For future work, we want to see if k-means clustering shows that car ownership is more likely for certain signs.