Feature Engineering

Ashrith Barthur, PhD

Engineered Features are Objective

- An investigator looks for large value transactions in the last 30 days.
- You code the average size of transactions for every week for the last 4 weeks.

Process of Engineering Features are Subjective

- An investigator looks for large value transactions in the last 30 days.
- You code the average size of transactions for every week for the last 4 weeks.
- While I code the average size of transactions for every 15 days for 2 fortnights, dissimilar to your feature.

Feature Engineering - A Subjective Process

- Comparing features is futile. Many roads, same destination.
- Targeting a good outcome is a worthy cause - high accuracy, recall, precision.
- Feature engineering is not a competitive process. It is instead a collaborative process.
- Utilize all the insights available in your organisation, peer group to make your features
-

Feature Engineering - A Subjective Process

- Do not try to encompass the universe of features that are relevant for your problem.
- **Follow the 80, 20 rule. If you have enough features to get you really close to the accuracy, try and tweak them instead of creating more features.**
- **Remember we are not looking for a theoretically amazing model, we are looking for valuable business insights!**

Engineering a Function of Time Feature.

- Data we will see today consists of one predictor.
- Contains login time of users on a social network.
- Data spans across 14 days

Engineering a Function of Time Feature.

time	user
1483338960	1
1483341840	2
1483385160	1
1483387800	2
1483424340	1
1483425240	1
1483427220	1
1483428240	2
1483429260	2

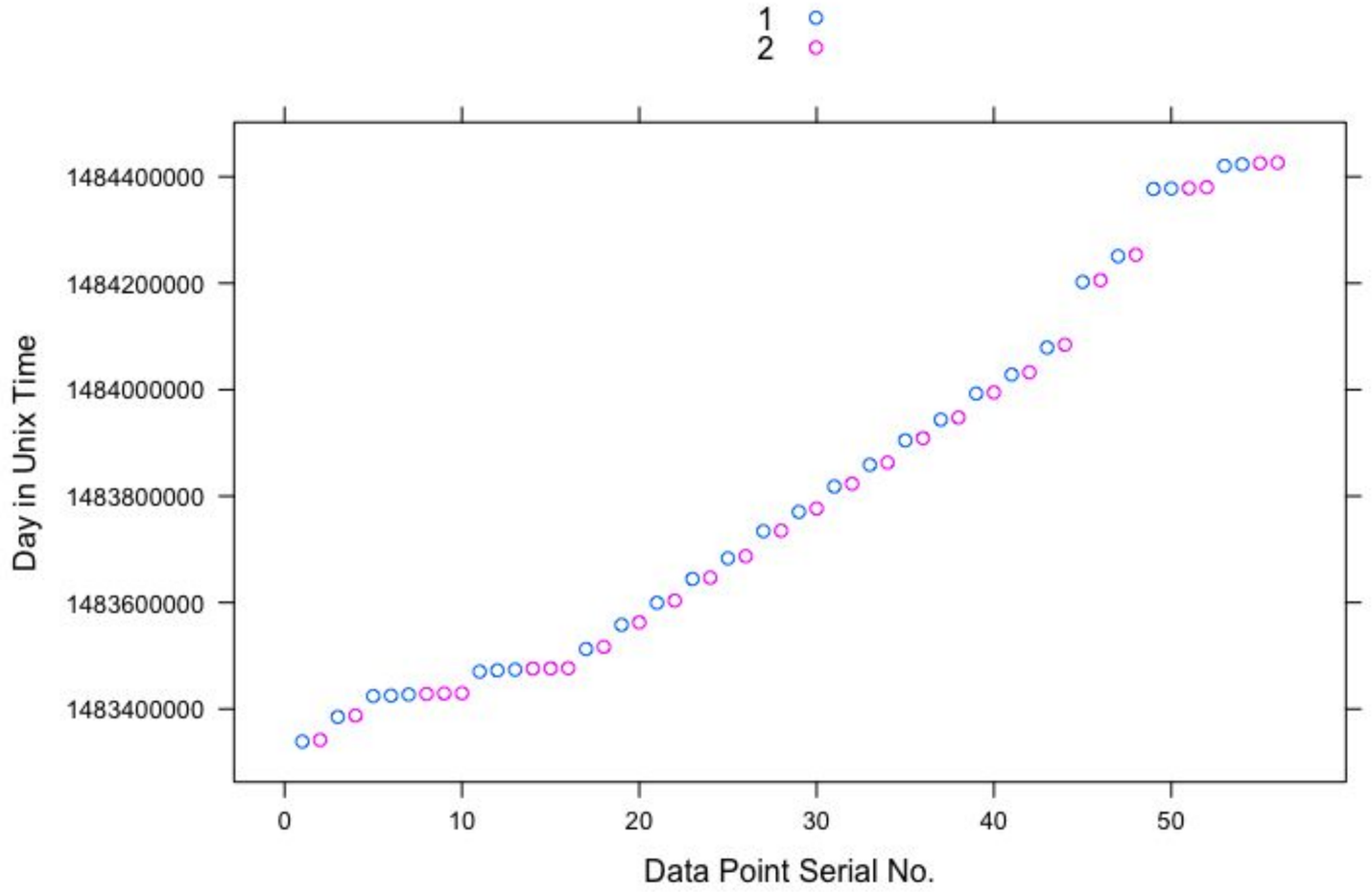
Engineering a Function of Time Feature.

- Single predictor column (time)
- Response column is multiclass, but in this case pure binary.

Engineering a Function of Time Feature.

- Specific example to keep things simple.
- Hopefully highlight effective feature engineering on using machine learning.

Visual Analysis of Distribution



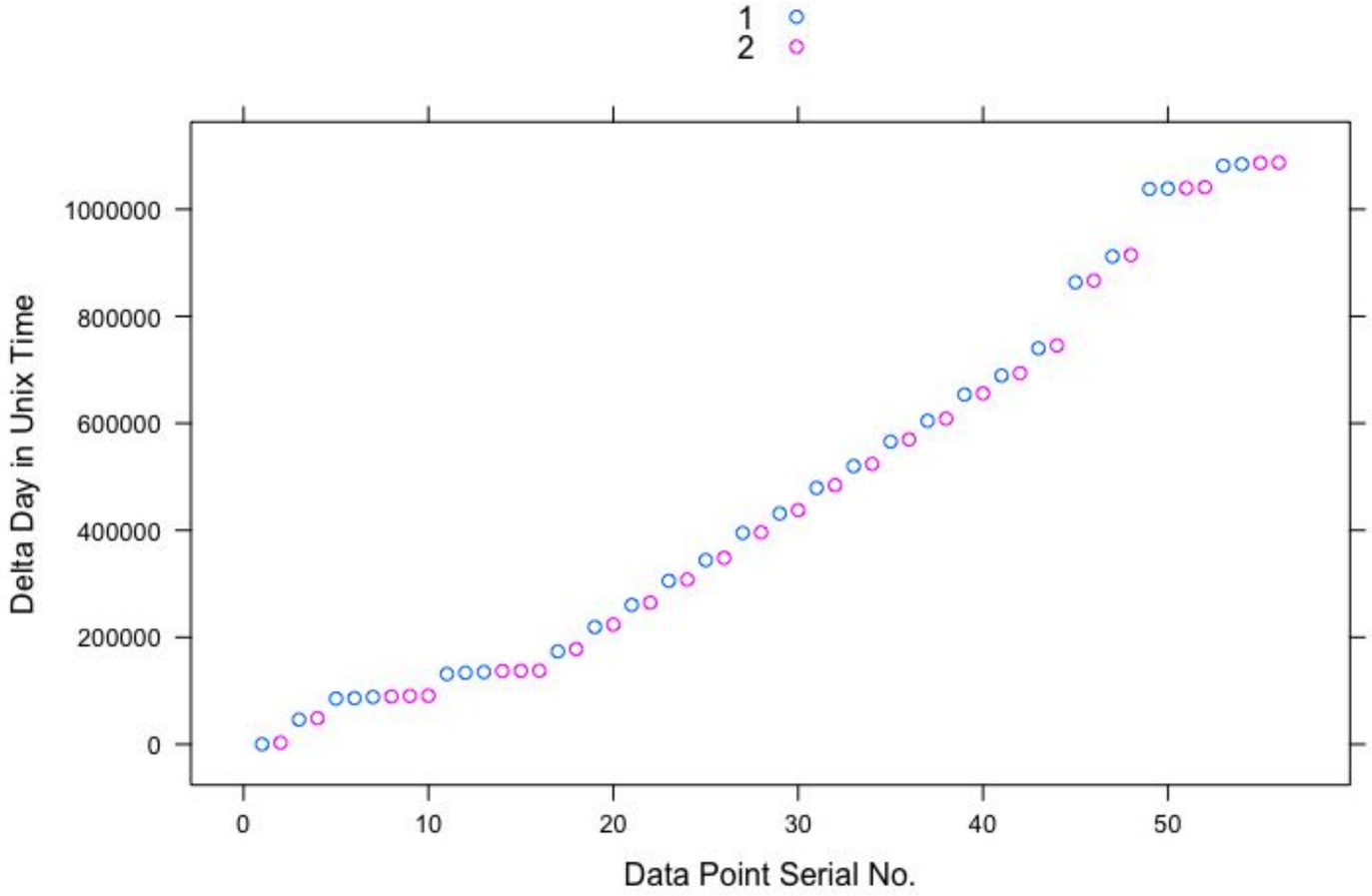
Visual Analysis of Distribution

- Monotonic (good proof that it could be seasonal)
- But what kind of seasonality? Weekly? Daily? Hourly?
- Two users show similar frequency.
- Let us run a machine learning model on it.

Outcome of Machine Learning model

- The Machine Learning model is not naturally inclined to identify seasonality.
- Low AUC shows it is missing the behaviour.
- Do we see autocorrelation? Prime for autoregression?

Visual Analysis of Distribution- Delta



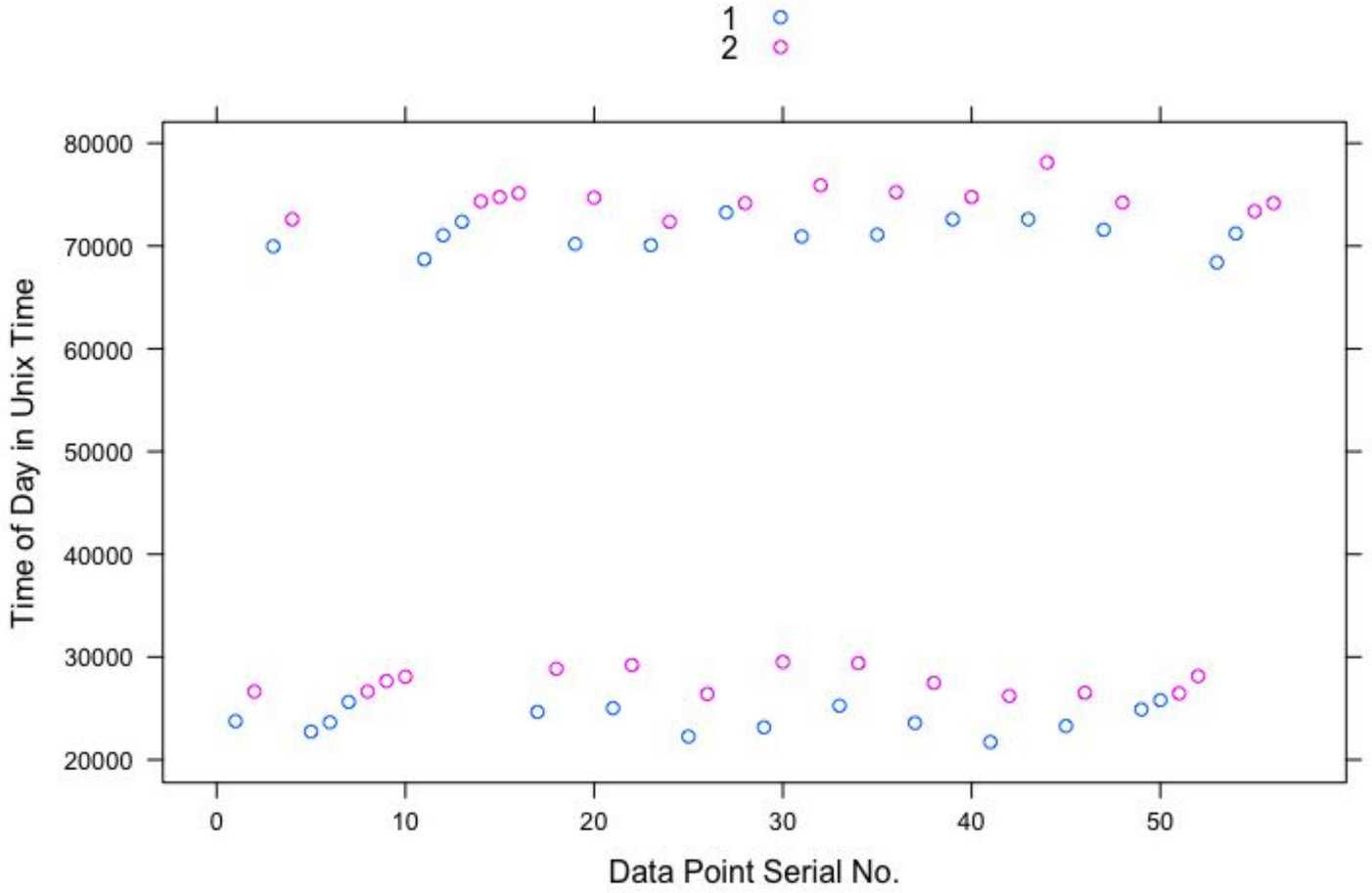
Visual Analysis of Distribution - Delta

- Not much difference.
- Just moved the intercept.
- Could there be a non-standard season? Across a day?

Engineering

- Let’s decompose the time data to a daily cycle.

Visual Analysis of Distribution-Decomposed



Visual Analysis of Distribution - Decomposed

- We see seasonal behaviour across a day.
- There are two periods of behaviour,
- 20000th - 30000th second of the day. 6 AM - 9:50 AM
- And 70000th - 80000th second of the day 7 PM to 10 PM

Engineering

- Let’s create a new feature with daily period.
- This is a new signal for the model
- Now let’s run the model again.

Outcome of Machine Learning model

- Compared to the previous model the machine learning model likes the seasonality.
- Relative better AUC.

Conclusion

- Machine Learning model is not smart to sight signals all the time.
- It needs some help.
- The help is feature engineering.
- Great features help the machines greatly!
- Features are Subjective!

Thank You

Questions?