CTR\_Prediction

|  |  |
| --- | --- |
| **Summary** | In this codelab, we demonstrate approach for Click-through Rate Prediction |
| **URL** | <https://codelabs-preview.appspot.com/?file_id=1DvTrj0tr7b4ekvTKgbTmAWQnb1oyPjccQ6-ebZR8F-w#0> |
| **Category** | Data Modeling |
| **Environment** | web, kiosk, |
| **Status** | Published |
|  |  |
| **Author** | Dingling Ge, Zefan Feng |
|  |  |
|  |  |

[Introduction](#_ok7k5uux6)

[Summary of the project and data](#_x0me2ipv51gl)

[Instructions](#_21yzqg98x7h6)

[DataSet](#_bwo0af2iwk90)

[Evaluation](#_bg3v8e57k1up)

[Data Preparation](#_76816rngv7is)

[Subset](#_1jjp7kh3dm0h)

[Hashing](#_rz6zrveltoh)

[LR](#_j40yxrbx796n)

[POLY](#_645pau9m1ve3)

[FM](#_ryousjrfhc1)

[FFM](#_pzs5zhk3gfjx)

[FFM CTR Prediction](#_bwox961q1a30)

[Conclusion](#_jdmzwkcs72o)

[Something else](#_zid1uwvx8dga)

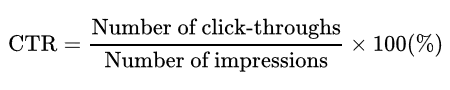
[Appendix](#_ot43tirxvs6w)

# Introduction

**Last Updated:** 2019-11-06

## **Summary of the project and data**

In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance. As a result, click prediction systems are essential and widely used for sponsored search and real-time bidding.



The data we use is from [Kaggle](https://www.kaggle.com/c/avazu-ctr-prediction/data). For the data file:

File descriptions

train - Training set. 10 days of click-through data, ordered chronologically. Non-clicks and clicks are subsampled according to different strategies.

test - Test set. 1 day of ads to for testing your model predictions.

sampleSubmission.csv - Sample submission file in the correct format, corresponds to the All-0.5 Benchmark.

## **Instructions**

* Know about dataset
* Evaluation
* Data Preparation

1. Subset
2. Hashing

* LR Model
* Poly Model
* FM Model
* FFM Model
* FFM CTR Prediction

# DataSet

Duration: 5:00

Our customer data comes from the [Kaggle](https://www.kaggle.com/) ,It’s easy to use, and free for non-commercial projects.

The dataset has the following columns:

Data fields

id: ad identifier

click: 0/1 for non-click/click

hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.

C1 -- anonymized categorical variable

banner\_pos

site\_id

site\_domain

site\_category

app\_id

app\_domain

app\_category

device\_id

device\_ip

device\_model

device\_type

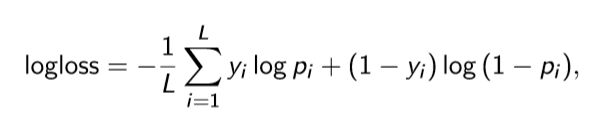
device\_conn\_type

C14-C21 -- anonymized categorical variables

# Evaluation

Duration: 5:00

Logarithmic loss is used in the model evaluation:



where L is the number of instances, yi ∈{0,1} is the label of the ith instance, and pi is the probability that the ith instance is clicked.

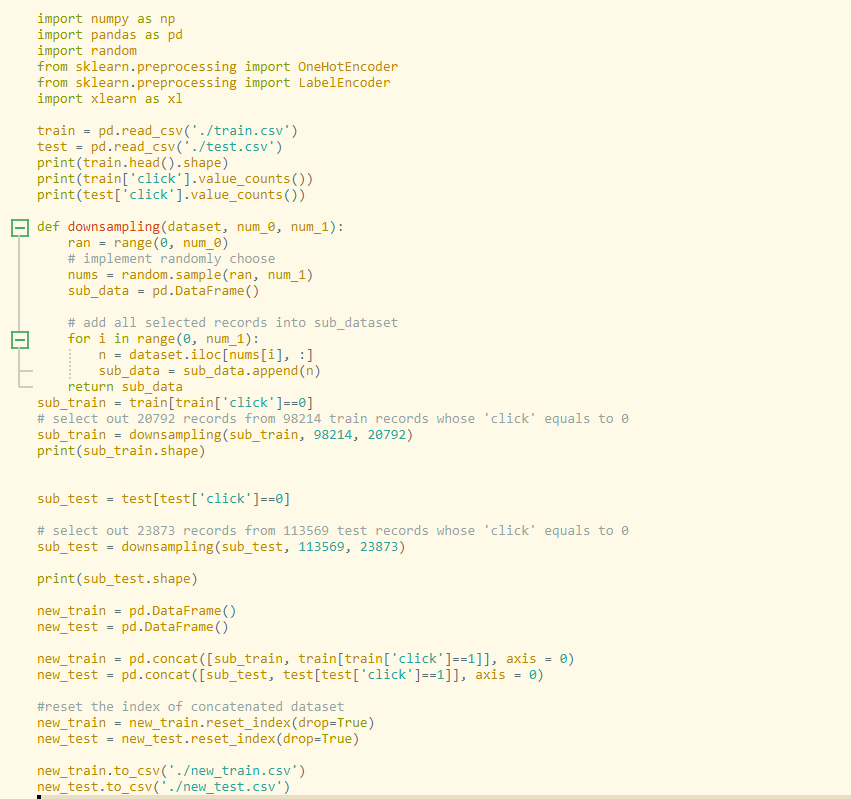
# Data Preparation

Duration: 20:00

## **Subset**

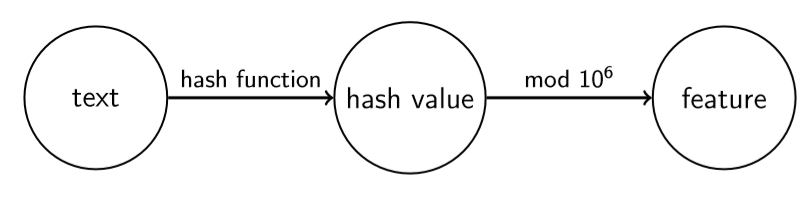
Instead of using the whole dataset, in this project we ﬁnd splitting data into small parts works better than directly using the entire dataset.

The proportion of positive and negative samples in the original data set is about 5:1, so we created a balanced test\_CSV which has 20792 records from 98214 ‘click’(which is 0)records and a balanced train\_CSV which has 23873 records from 113569 ‘click’(which is 0) to train the model.



## **Hashing**

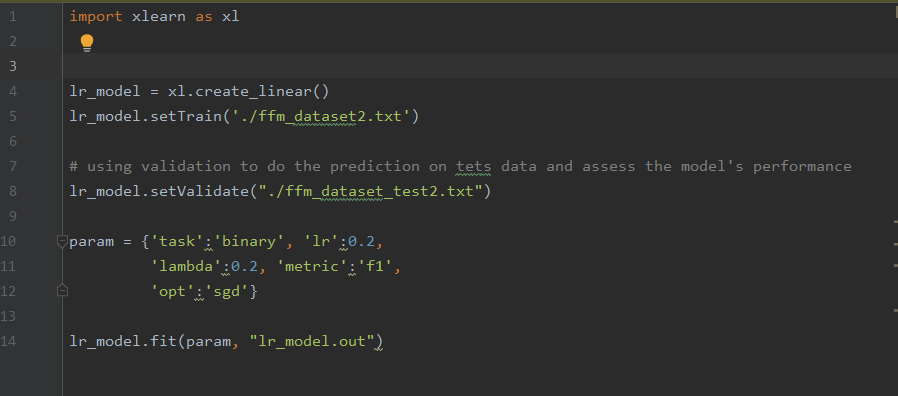
We use hashing trick to transform text features.



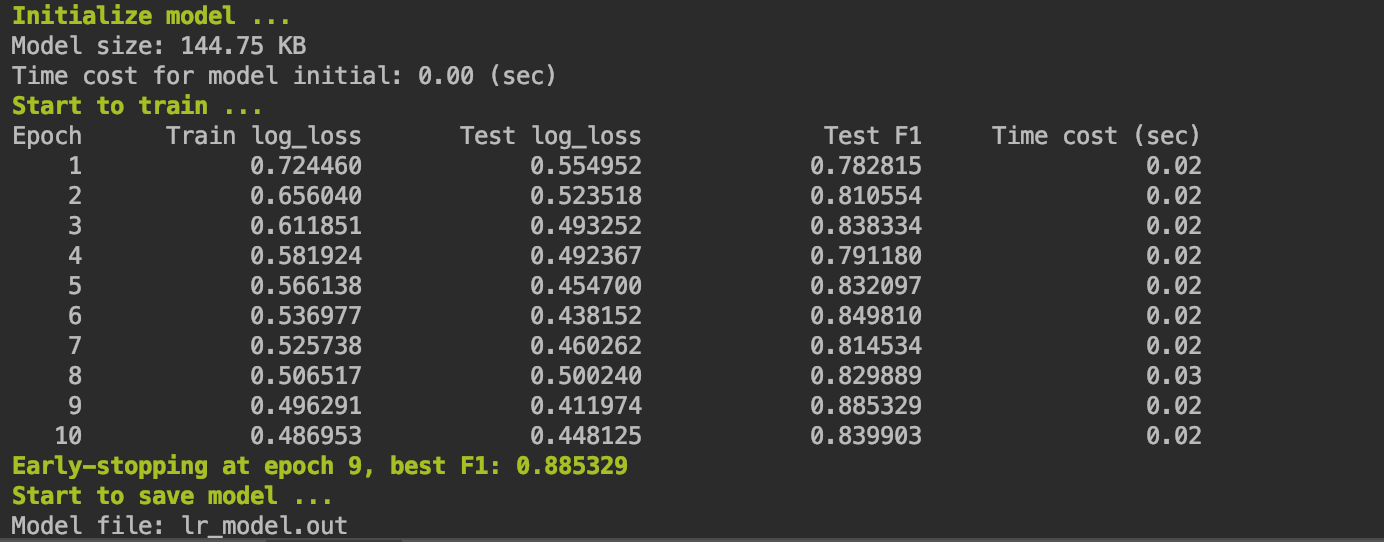
# LR

Duration: 10:00

Code:



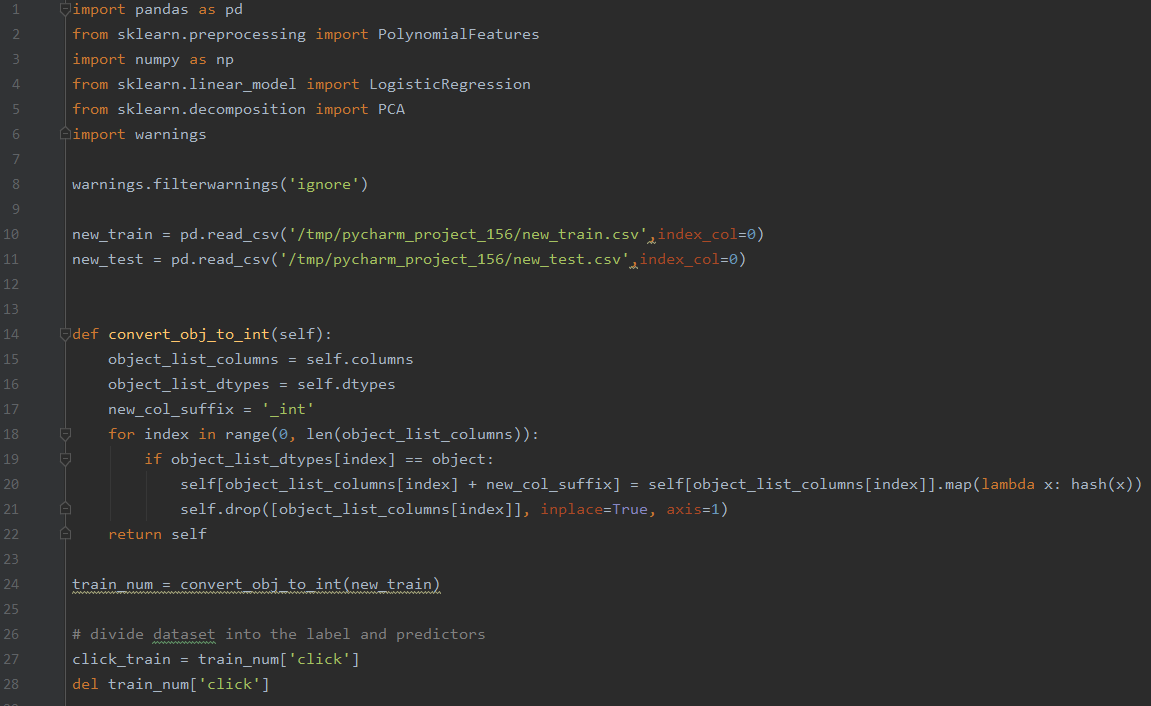
Performance:



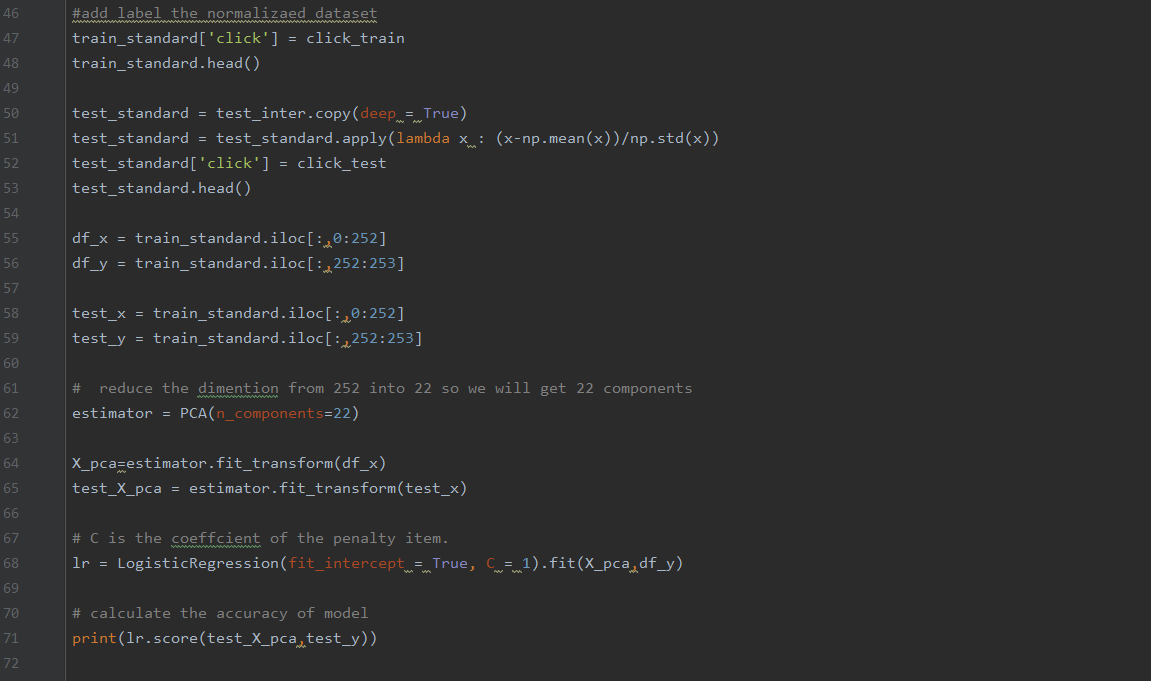
# POLY

Duration: 10:00

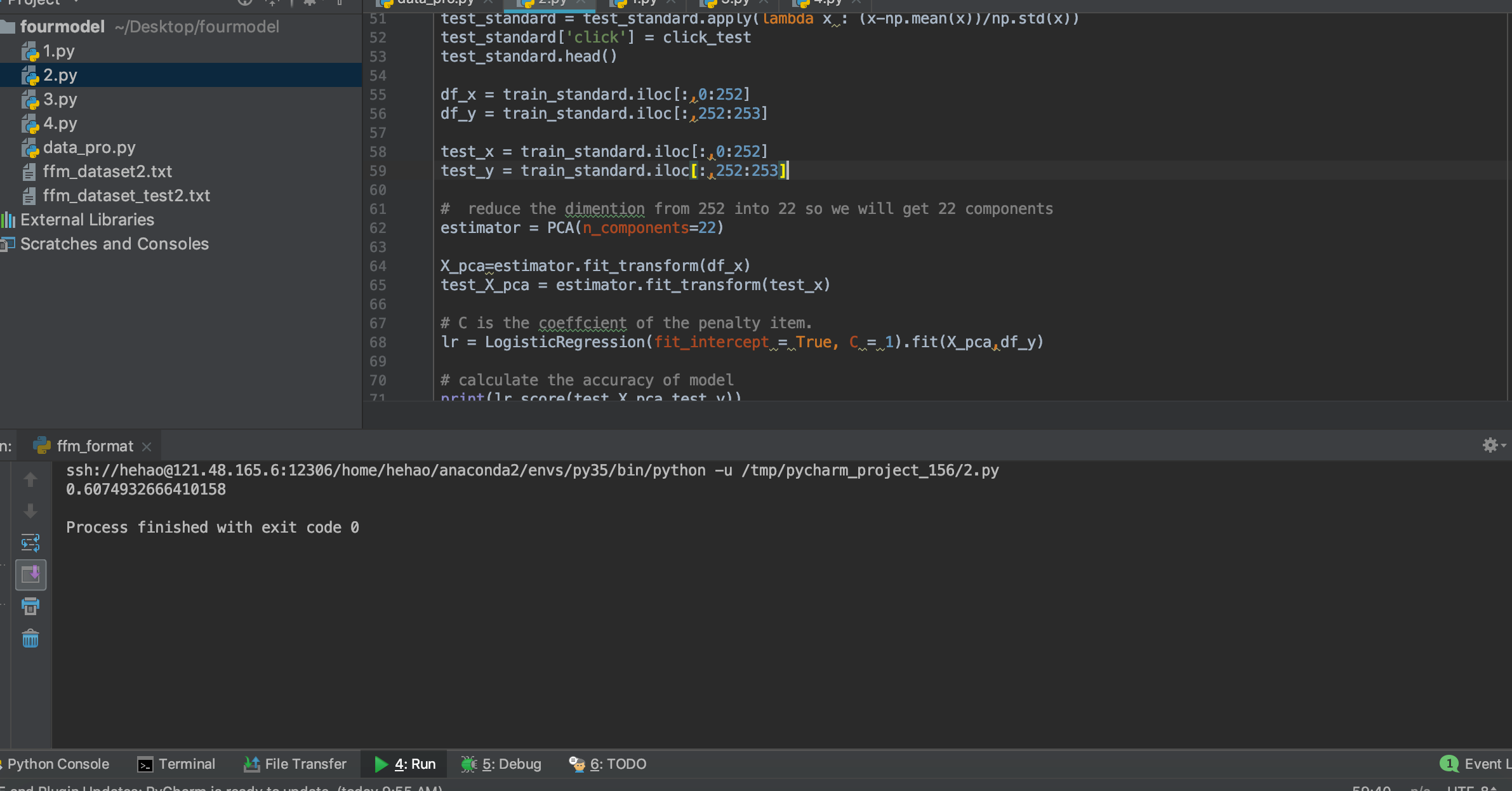
Code:







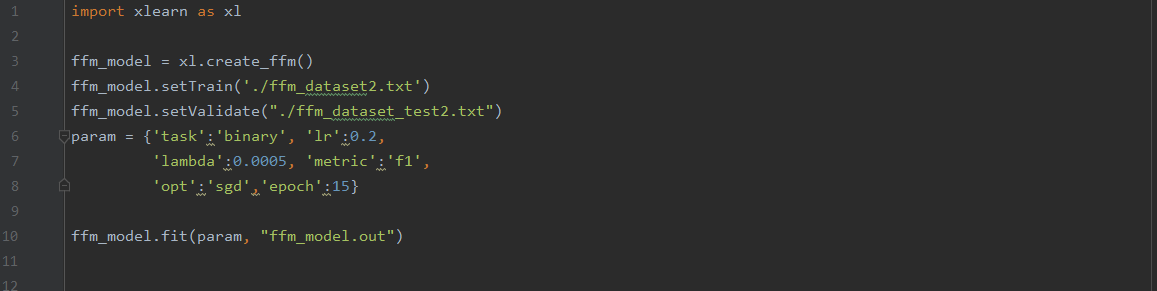
Performance:



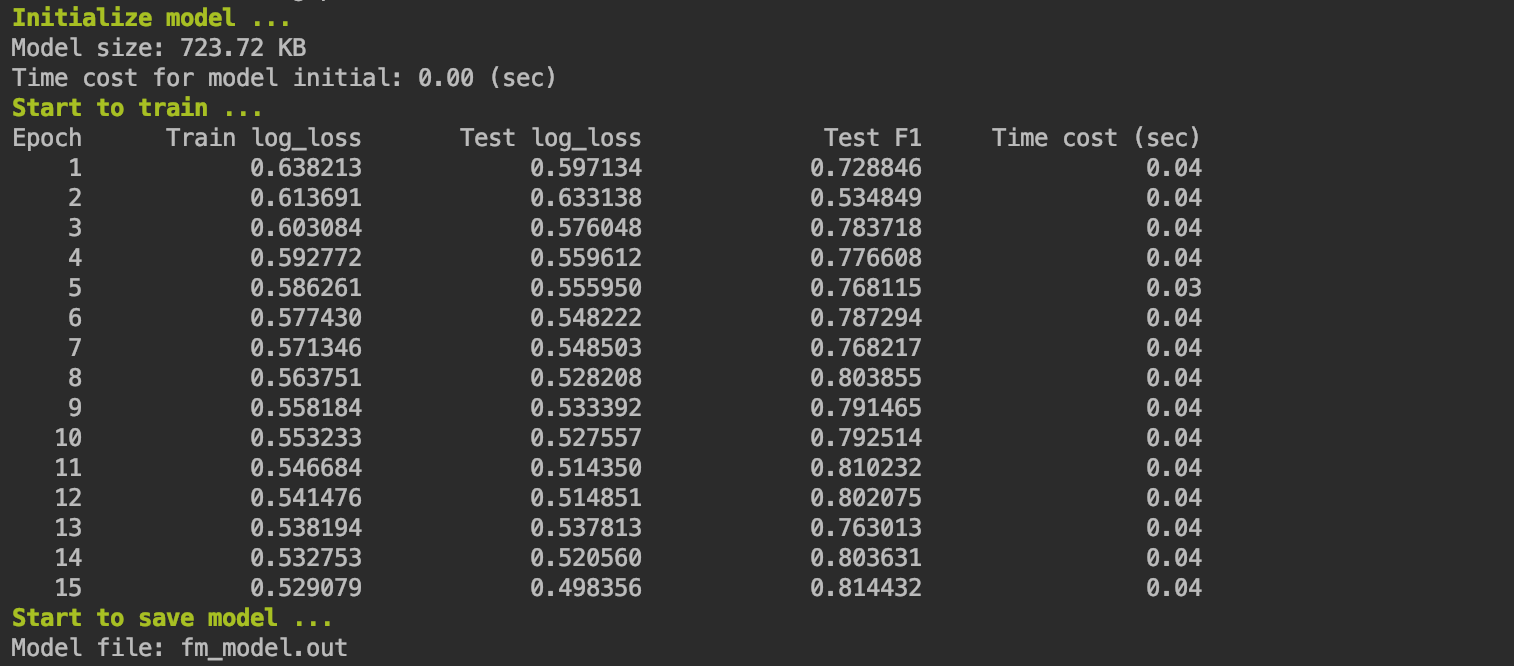
# FM

Duration: 10:00

Code:



Performance:



# FFM

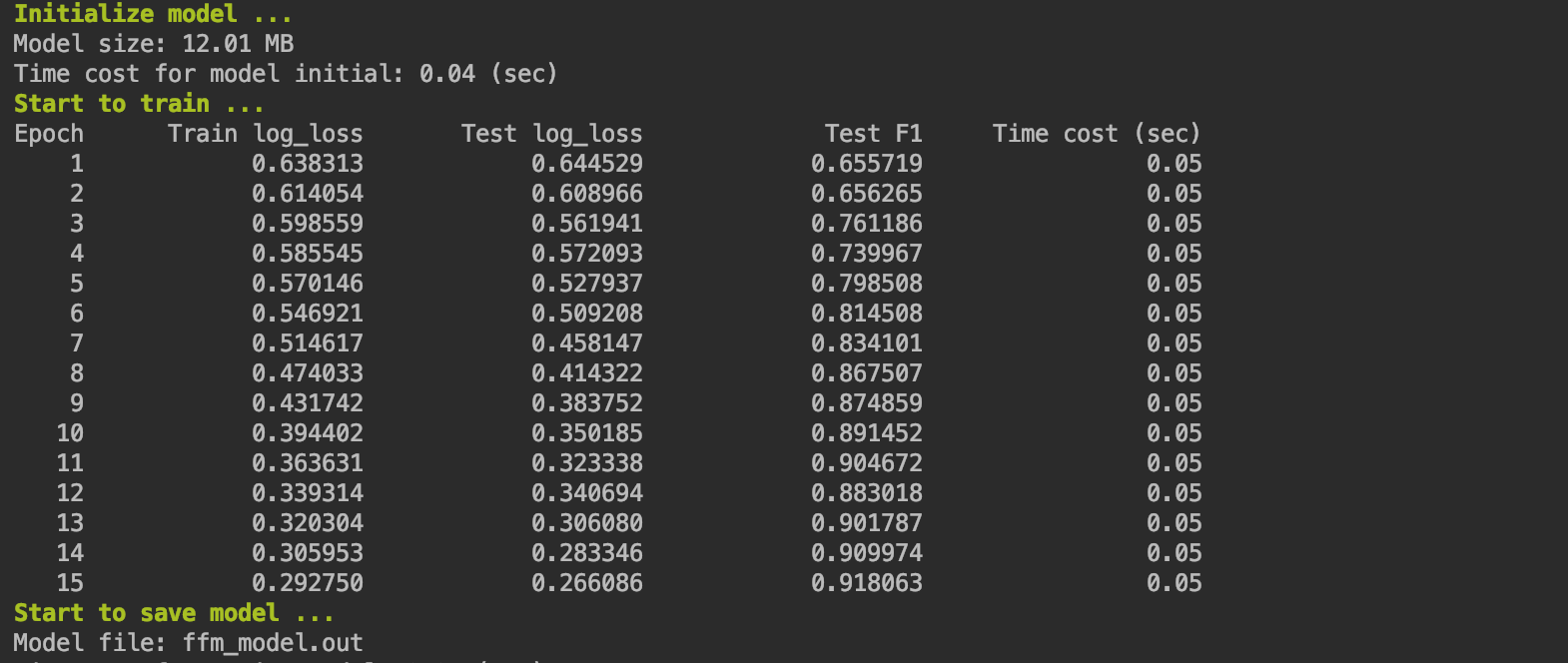
Duration: 10:00

Code:





Performance:

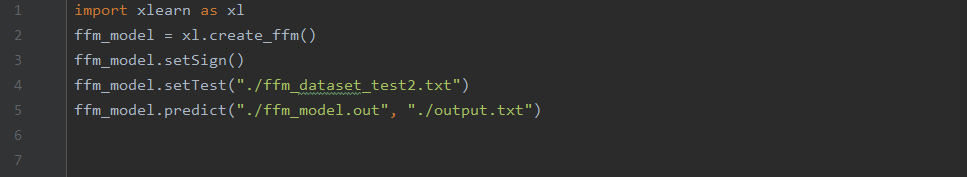


# FFM CTR Prediction

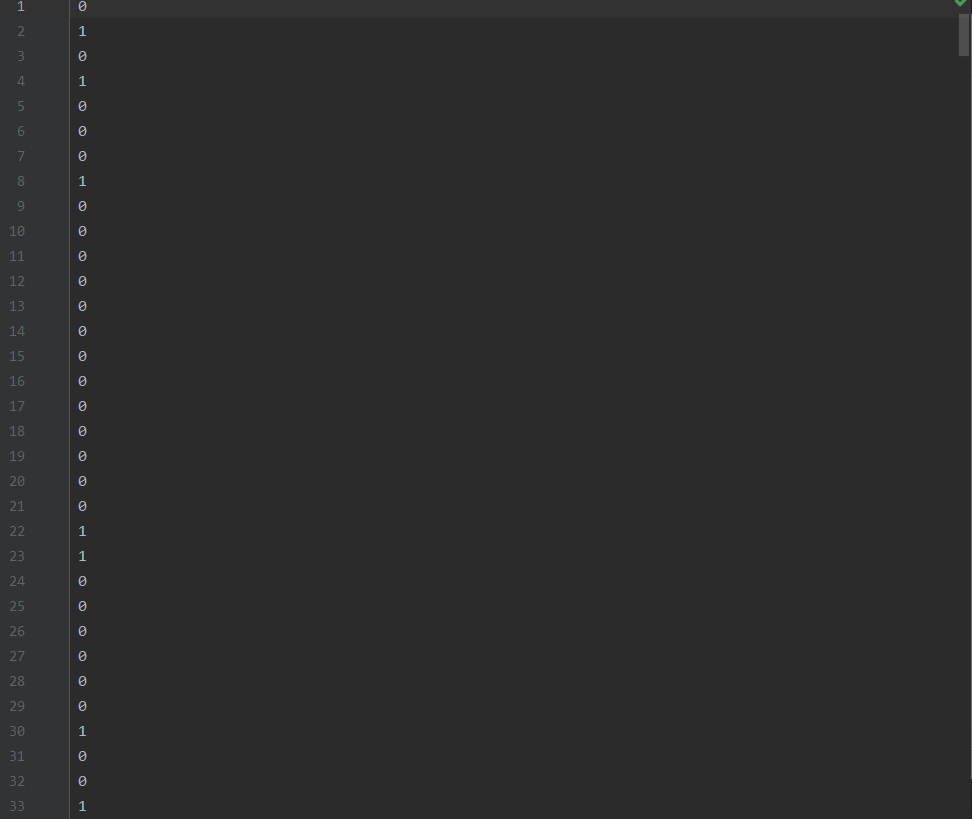
Duration: 10:00

We use ffm\_dataset\_test2.txt as input, and ffm.out as our model to make CTR predictions.

Code:



Output:



# Conclusion

Duration: 1:30

FFM is a better Model used to predict CTR based on the performance of the 4 models.

## **Something else**

LogLoss is a general method to evaluate models.

# Appendix

Duration: 3:00

You can look up the related analysis in [Kaggle](https://www.kaggle.com/c/avazu-ctr-prediction)

You can look up the detailed code on [Github](https://github.com/DinglingGe/Assignment-2)