Данные

Сайты, посещенные пользователями
(url_domain_train - 2046869
 url_domain_test - 613388)

- Titles посещенных сайтов

(title_unify_train - 5850389
 title_unify_test - 2303479)

url_domain_train

0	0000000151004FF4ADD746DA10685A01	afisha.ru	2
1	0000000151004FF4ADD746DA10685A01	aif.ru	1
2	0000000151004FF4ADD746DA10685A01	aimfar.solution.weborama.fr	1
3	0000000151004FF4ADD746DA10685A01	alkotest.ru	1
4	0000000151004FF4ADD746DA10685A01	aptekamos.ru	1
5	0000000151004FF4ADD746DA10685A01	aquavivo.ru	1
6	0000000151004FF4ADD746DA10685A01	arco-iris.ru	1
7	0000000151004FF4ADD746DA10685A01	autorambler.ru	3
8	0000000151004FF4ADD746DA10685A01	avtoprofi.ru	1
9	0000000151004FF4ADD746DA10685A01	championat.com	1

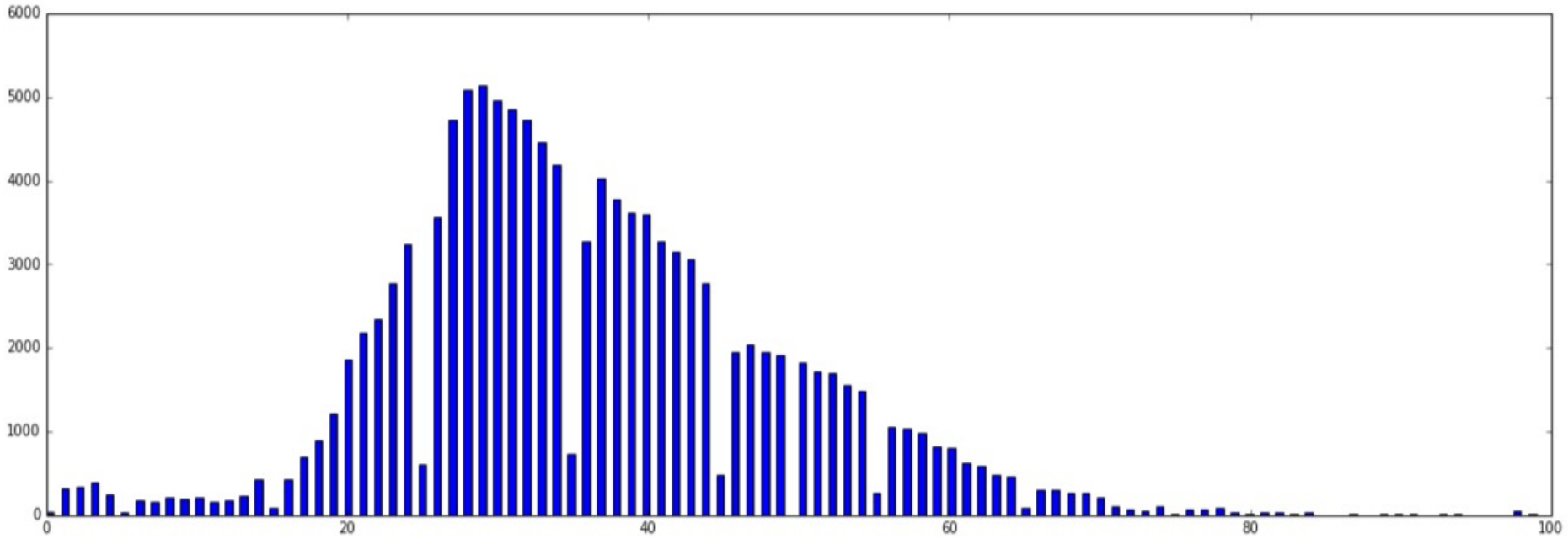
title_unify_train

	title	id	age
0	[бесплатный надёжный почта рамблер электронный...	000000013CB5719C0000A2C90002C101	53
1	[24-х 34-х до договор неделя новость предложит...	00000001442BE24000001B7D00F50801	48
2	[авто бош контакт королёв сервис, авто бош кор...	00000001448580F800003F1B31FB0901	28
3	[ua втрать війни донбасі за на новини озвучить...	0000000145BDB2FF000157971645E901	44
4	[black walnut грецкий орех чёрный, inmoment ru...	000000014602771F0000DB9359714C01	48
5	[апрель год день март месяц на неделя от погод...	0000000147B2D6F311DB5C4201B7FB01	36
6	[rankw ru tovar xoz xoz-tovar анализ доход ком...	0000000147C68954150168D701A8B801	33
7	[возможность госзакупка консультант малое плюс...	0000000147EB76D738CD80750C879701	41
8	[13d билет заказать имя купить на спектакль те...	00000001482AAFB69FA5228008AC2A01	51
9	[1-метр velol александр альберто арно велоспор...	0000000148390BB56A6B22BB178D3901	32
10	[1vitali1 ii мафра от порту санинга токио фк ц...	00000001487DAF8D69CD43E416D6AD01	29
11	[caloriz ru www диета курица модный на похуден...	0000000148AC192341E3BDAD0B95DE01	36
12	[plc powerlin адаптер, агентство азербайджан и...	0000000148C2B61B70F8651309287201	35
13	[матч новость премьер-лига расписание россия р...	0000000148DB999352D9CAF309511101	37
14	[window бесплатно для живой на обои рабочий ск...	0000000148EB77A2435E76A711E38B01	37

Целевая переменная - возраст

In [16]:
```
%pylab inline
pylab.figure(figsize=(20, 6))
plt.hist(y, bins=200)
plt.show()
```

```
Populating the interactive namespace from numpy and matplotlib
```

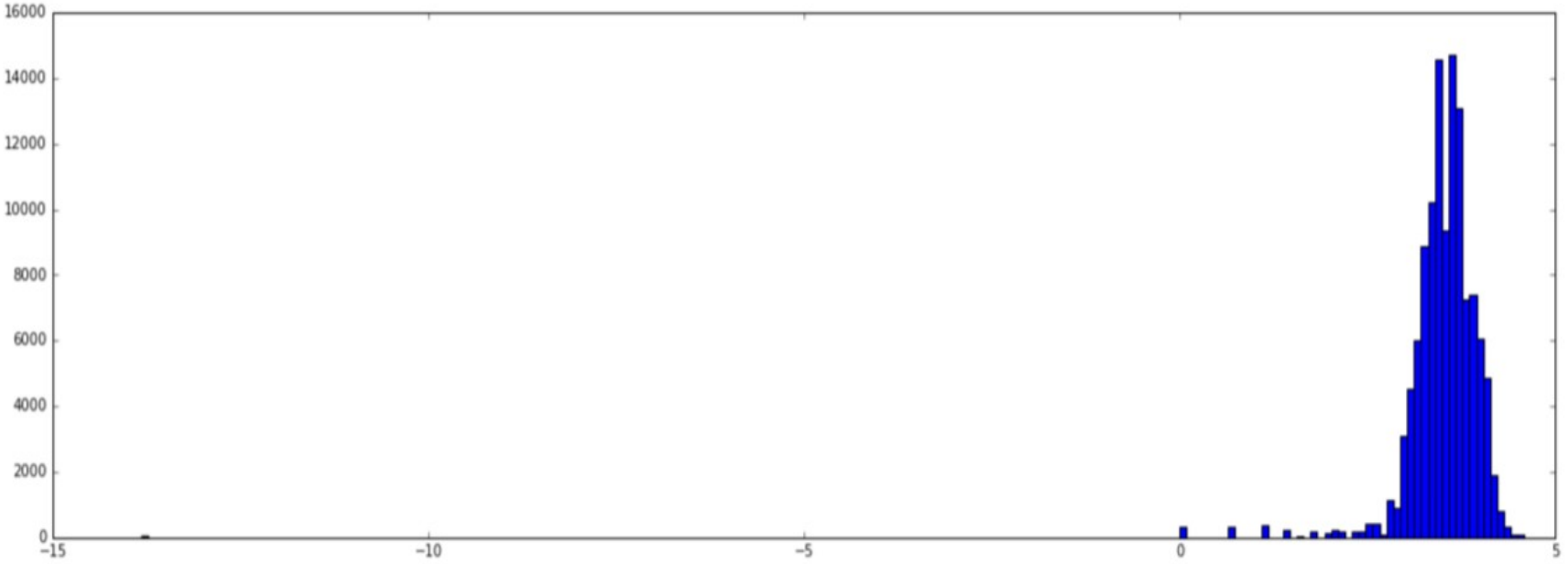


Логарифм от целевой переменной

```
In [30]: y_log = np.log(y +0.000001)
```

```
In [31]: %pylab inline
         pylab.figure(figsize=(20, 6))
         plt.hist(y_log, bins=200)
         plt.show()
```

```
Populating the interactive namespace from numpy and matplotlib
```



Представление данных

- HashingVectorizer
- (Число ячеек подбирается вручную)
- Отдельно для url(1800) и title(3500)

Модели над url'ами

- Линейная регрессия
- Линейная регрессия над tfidf
- Бустинг
- Бустинг над tfidf

Линейные

Линейная регрессия

In [12]:
```
reg = LinearRegression()
reg.fit(train_data, train_labels)
```

Out[12]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

In [13]:
```
linear_pred = reg.predict(test_data)
```

In [14]:
```
rmse(linear_pred, test_labels)
```

Out[14]:
```
11.812334918128547
```

Линейная регрессия над tfidf

In [15]:
```
linear_tfidf = pipeline.Pipeline([('tfidf', feature_extraction.text.TfidfTransformer()),
                                  ('linear_model', linear_model.LinearRegression())])
linear_tfidf.fit(train_data, train_labels)
```

Out[15]:
```
Pipeline(steps=[('tfidf', TfidfTransformer(norm=u'l2', smooth_idf=True, sublinear_tf=False,
         use_idf=True)), ('linear_model', LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=Fals
e))])
```

In [16]:
```
pred_linear_tfidf = linear_tfidf.predict(test_data)
rmse(pred_linear_tfidf, test_labels)
```

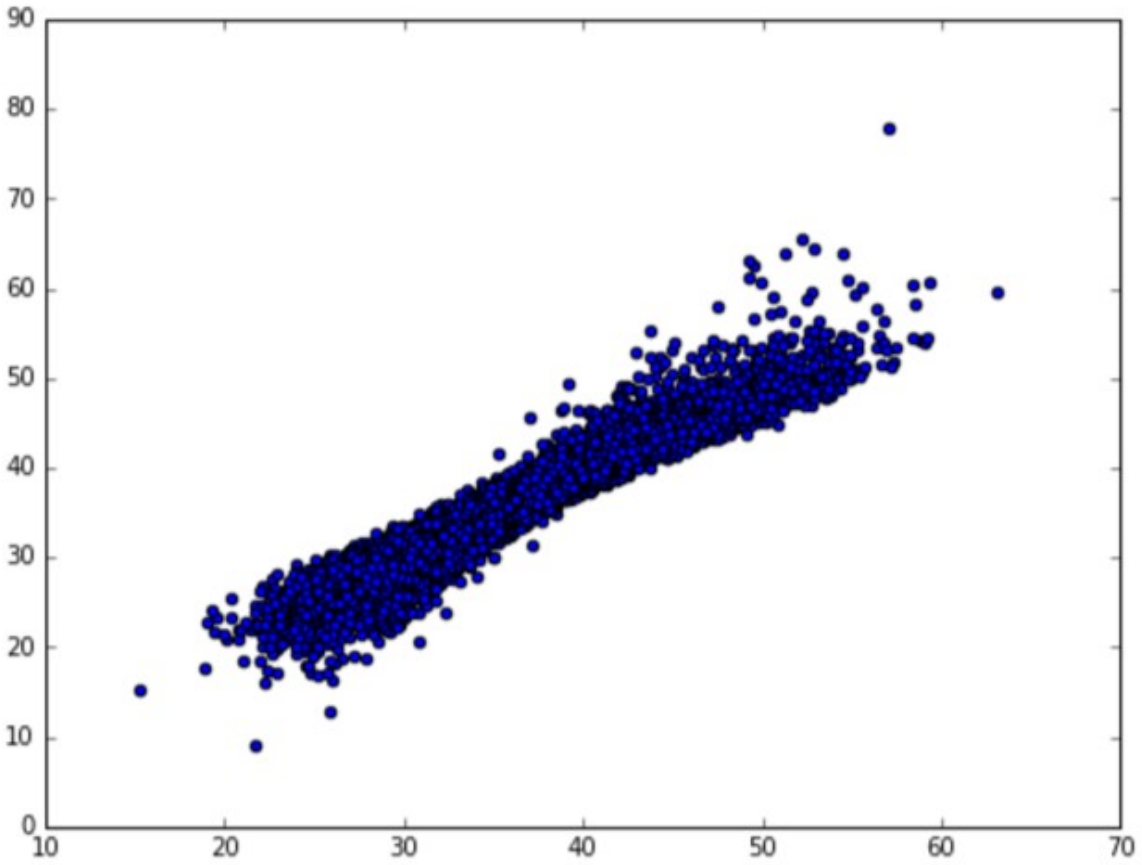
Out[16]:
```
11.748692764146533
```

По осям — разные предсказания

```
In [23]: %pylab inline
         pylab.figure(figsize=(8, 6))
         pylab.scatter(pred_linear_tfidf, linear_pred)
```

```
Populating the interactive namespace from numpy and matplotlib
```

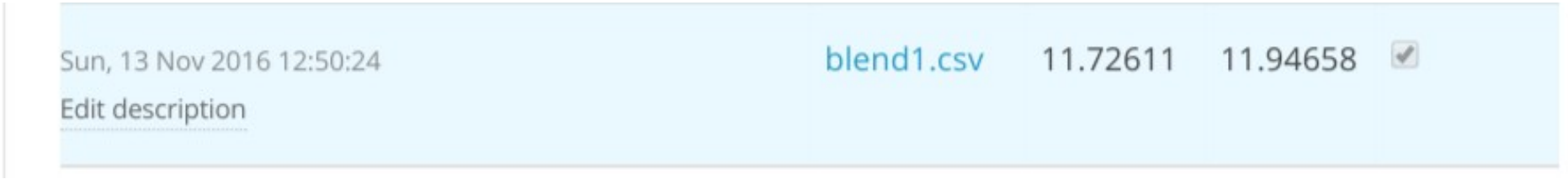
```
Out[23]: <matplotlib.collections.PathCollection at 0x7f7af5edff10>
```



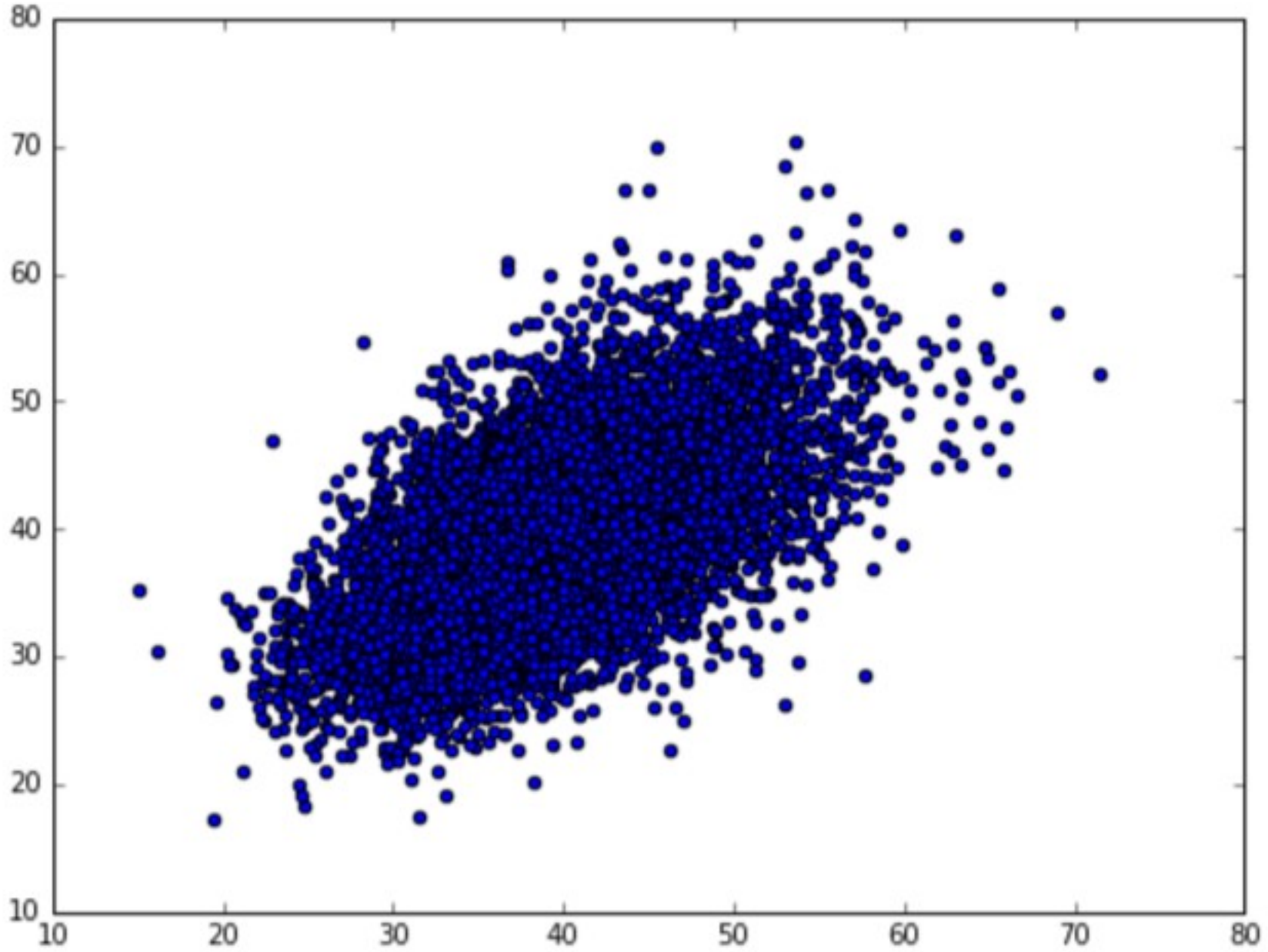
Аналогичные модели для title

- Линейная регрессия
- Линейная регрессия над tfidf
- Бустинг
- Бустинг над tfidf

Взял среднее от всех 6 моделей.



boost_title/boost_url



Среднее от бустинга над titles_tfidf и над url_tfidf