Данные

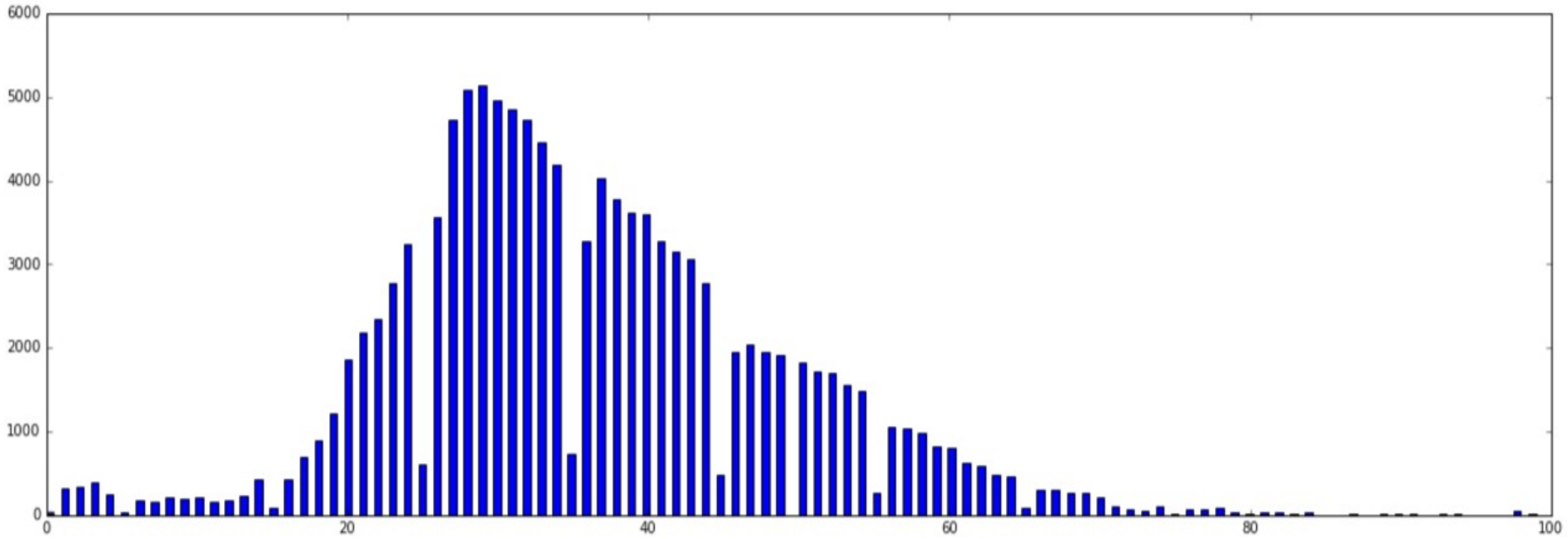
- Сайты, посещенные пользователями (url_domain_train)
- Titles посещенных сайтов

  (title_unify_train)

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
```

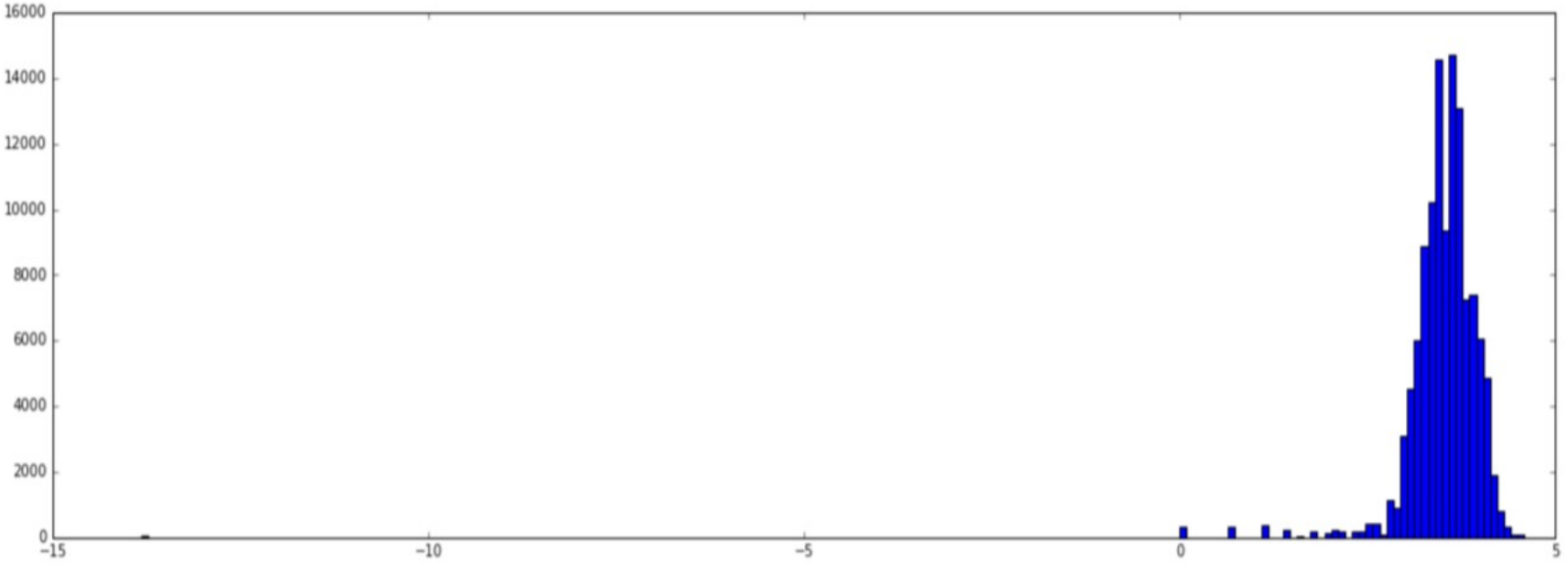


Логарифм от целевой переменной

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)
```

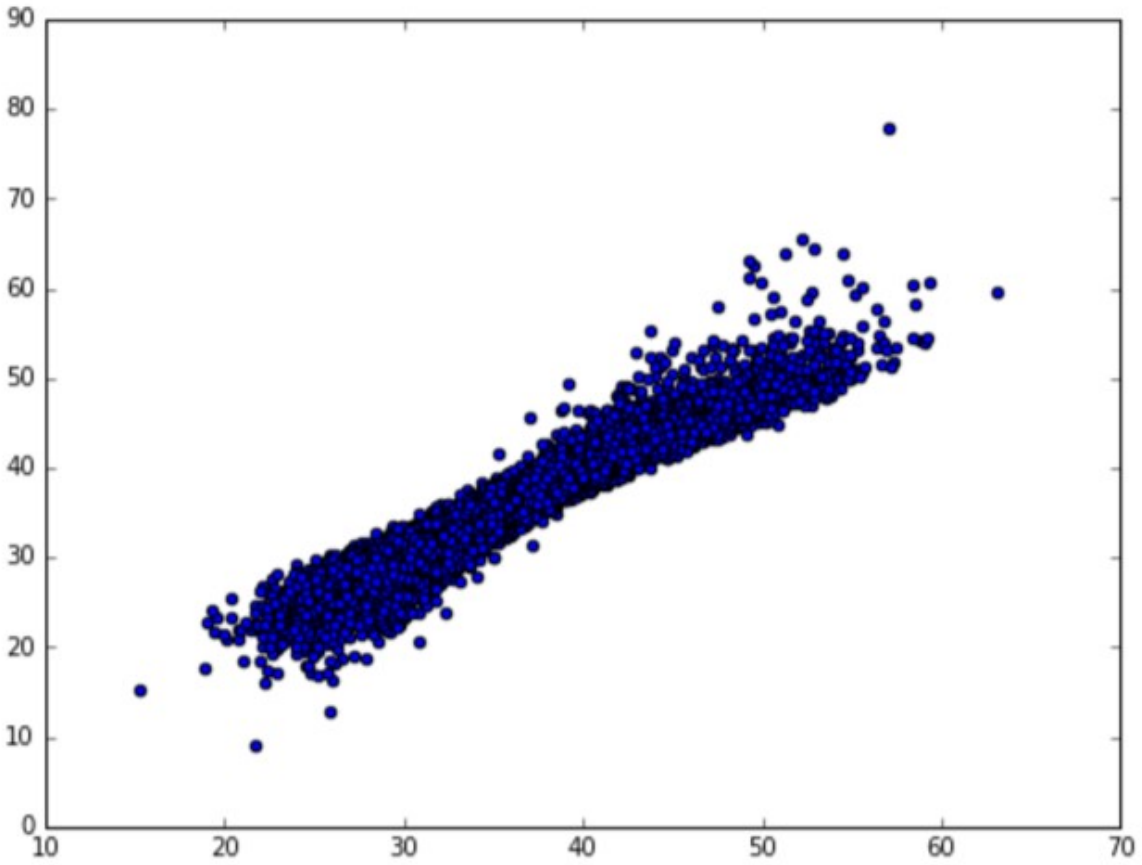
Out[16]:
```
11.748692764146533
```

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
```

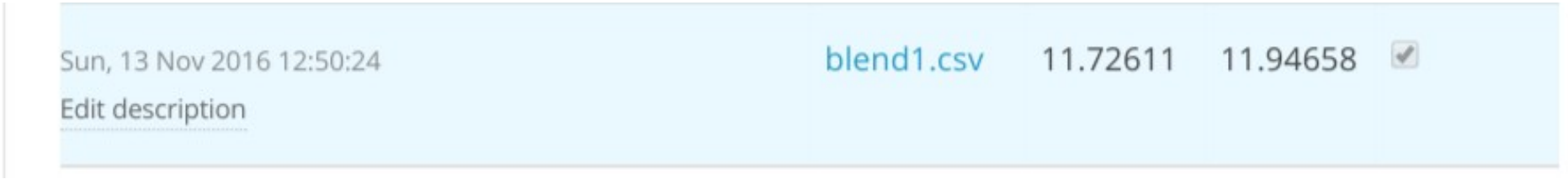
```
Out[23]: <matplotlib.collections.PathCollection at 0x7f7af5edff10>
```



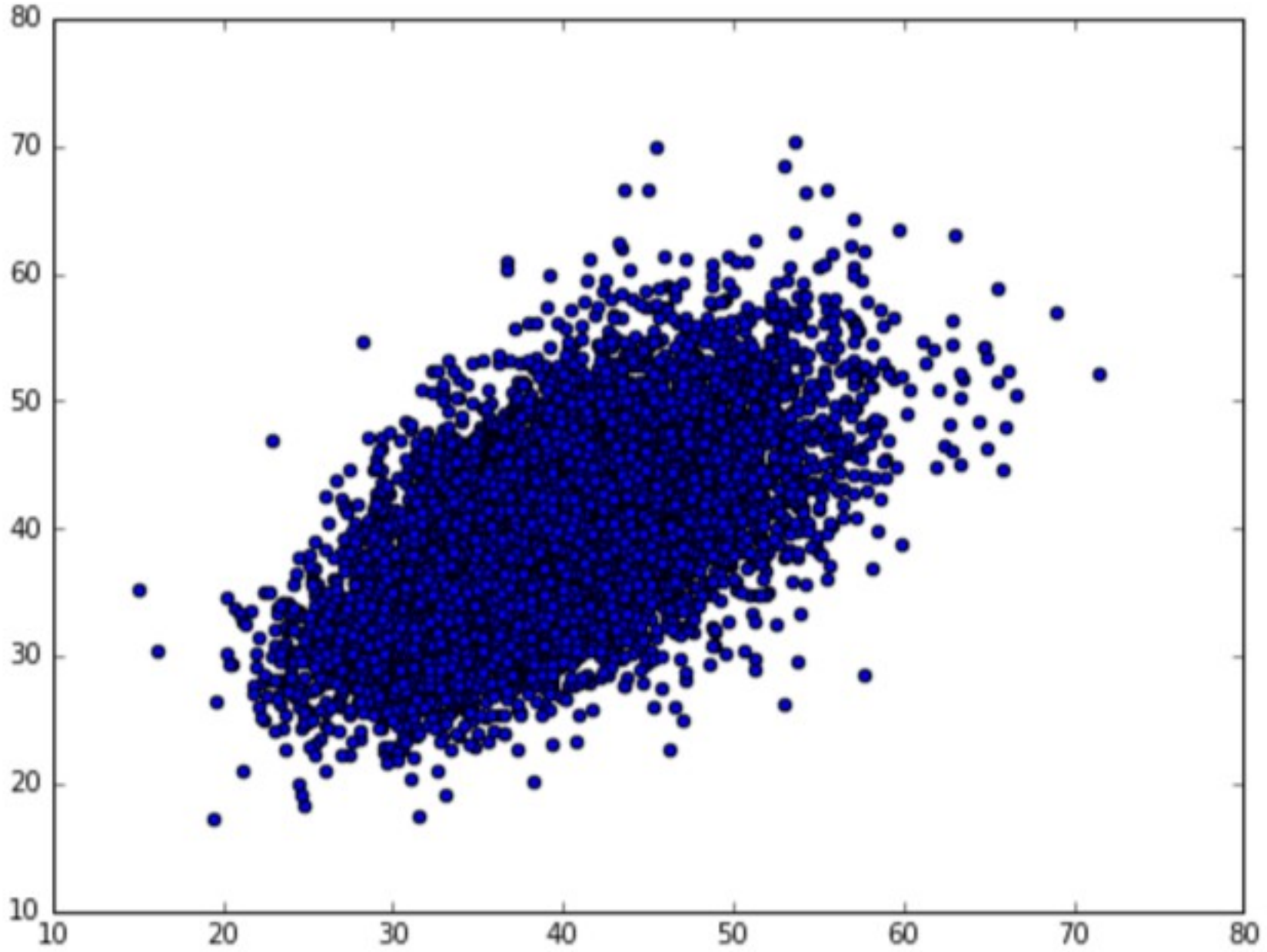
Аналогичные модели для title

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

Взял среднее от всей 6 моделей.



boost_title/boost_url



Среднее от бустинга над titles_tfidf и над url_tfidf