# Анализ информации по электронным торгам

Веревкина Е.Б., студентка кафедры информатики СПбГУ,
verevkinahelen@gmail.com

## Аннотация

Статья посвящена проблеме прогнозирования итоговых цен лотов, продаваемых на электронных торговых площадках. Описаный в статье подход позволил создать программный продукт, предоставляющий пользователю возможность получить быстрый и достаточно точный прогноз итоговой цены еще не реализованного лота. Опираясь на полученную информацию, пользователь может скорректировать свою стратегию участия в торгах и одержать победу.

## Введение

В условиях кризиса банкротство физических и юридических лиц стало частым явлением. При наличие у банкрота имущества, на которое может быть обращено взыскание, происходит его опись, оценка и составление плана продажи. Имущество должника продается на торгах, которые проводятся на различных электронных торговых площадках (ЭТП). Электронные торги позволяют увеличить аудиторию потенциальных покупателей и быстро продать имущество должника. Цена имущества на таких торгах часто бывает крайне низкой, она может достигать 20% от рыночной. Именно поэтому торги на ЭТП интересны как предпринимателям, так и частным лицам.

Анализ данных по завершившимся торгам представляет большой интерес с точки зрения прогнозирования результатов будущих торгов.

Основной задачей настоящей работы является нахождение факторов, оказывающих наибольшее влияние на изменение цены лота в ходе торгов, с целью прогнозирования итоговой цены лота на основе модели линейной регрессии. Для выявления закономерностей был проанализирован набор данных по электронным торгам, на которых продавались автотранспортные средства.

Приведенное в статье программное решение призвано расширить функционал сервиса bankrot-spy.ru[1].

[1] bankrot-spy.ru — сервис, который объединяет информацию о торгах, проводимых на различных ЭТП, предоставляет статистику и помощь в торгах.

## Теоретическое обоснование

Метод линейной регрессии является распространенным способом прогнозирования зависимой величины на основе одной или нескольких независимых переменных.

В методе множественной линейной регрессии связь переменной $Y$ с переменными $X_1, ..., X_n$ задается с помощью линейной модели

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon,$$

где $\beta_0, \beta_1, ..., \beta_n$ — вещественные регрессионные коэффициенты, $\varepsilon$ — случайная величина, являющаяся ошибкой прогнозирования [2].

Регрессионные коэффициенты ищутся по обучающей выборке — набору данных о завершенных торгах. Таким образом, задача прогнозирования сводится к нахождению прямой, которая будет наилучшей аппроксимацией данных из обучающей выборки.

Основной задачей является прогнозирование финальной цены для еще незавершенных торгов, следовательно, именно финальная цена и будет единственным выходным параметром линейной регрессии. С целью выявления наилучшего с точки зрения качества прогноза набора входных параметров, было проведено сравнение коэффициентов корреляции между финальной ценой и другими параметрами лота, построено несколько регрессионных моделей с разными входными параметрами, рассмотрены их диаграммы рассеяния и коэффициенты детерминации.

Так как необходимым условием создания применимой на практике модели прогнозирования является наличие корректно определенных входных параметров регрессии как у реализованных, так и у нереализованных лотов, были выявлены следующие возможные регрессоры: начальная цена лота, регион, номер квартала, в котором был начат прием заявок.

Рассмотрим самую простую и очевидную зависимость — зависимость итоговой цены от начальной. Очевидно, именно эта зависимость послужит фундаментом для добавления прочих параметров, улучшающих прогноз.

## Зависимость финальной цены лота от начальной

Коэффициент корреляции характеризует меру линейной зависимости двух переменных. Рассмотрим коэффициент корреляции Пирсона

между начальной и финальной ценами лота, который задается формулой

$$r = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2 \sum\limits_{j=1}^{n}(Y_j - \bar{Y})^2}}$$

где $X_i$ — начальная цена лота, $Y_i$ — финальная цена лота, $\bar{X}$ и $\bar{Y}$ — выборочные средние, определяющиеся следующим образом

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i),$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(Y_i).$$

Коэффициент корреляции оказался равным 0,868, это позволяет сделать вывод, что между начальной и финальной ценами существует достаточно сильная линейная зависимость. Следовательно, будет уместно применить линейную регрессию для прогнозирования финальной цены на основе начальной.

Для оценки качества построенной модели линейной регрессии воспользуемся диаграммой рассеяния (см. Рис. 1). По диаграмме видно, что разброс между эталонными и вычисленными значениями невелик. Следовательно, данные хорошо укладываются в линейную модель, и на ее основании можно сделать прогноз на будущее.

Теперь рассмотрим коэффициент детерминации такой модели [2]. Коэффициент детерминации показывает, какая доля вариации зависимой переменной $Y$ учтена в модели и обусловлена влиянием на нее факторов, включенных в модель

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum\limits_{i=1}^{n}(Y_i - \bar{Y})^2},$$

где $Y_i$ — значения зависимой переменной, $\bar{Y}$ — среднее значение зависимой переменной, $\widehat{Y}_i$ — модельные значения, построенные в результате применения линейной регрессии.

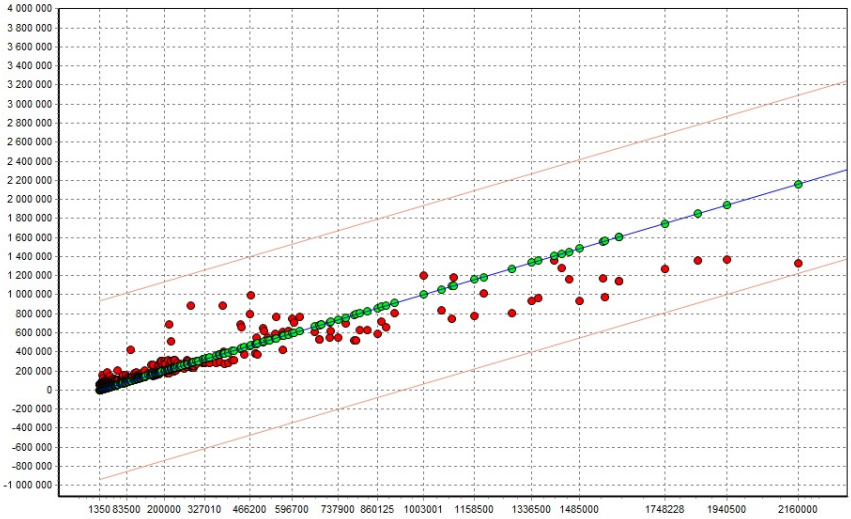


Рис. 1: Диаграмма рассеяния парной линейной регрессии

Коэффициент детерминации для модели линейной регрессии с одним входным параметром — начальной ценой, и одним выходным параметром — финальной ценой равняется 0,754.

Для улучшения этого показателя введем дополнительные входные параметры.

## Категориальные параметры регрессии

Значения категориальных признаков определяет факт принадлежности объекта к какой-либо категории. В рассматриваемом наборе данных категориальными признаками являются регион, квартал года, а также модель автомобиля, но модель не задается явно при публикации информации о торгах и извлекается корректно из названия лота только в 20% случаев, поэтому этот признак не является достаточно надежным для использования в итоговой модели.

Часто категориальные признаки представлены строковыми значениями, поэтому, прежде чем использовать их в регрессионной модели, нужно провести нормализацию, то есть преобразовать к виду наиболее подходящему для обработки алгоритмом. В случае линейной регрессии

это числовой тип.

Существует несколько способов преобразования категориальных данных. Рассмотрим их на примере регионов.

- Кодирование уникальным значением. Сопоставим каждому региону целое число (см. Таблицу 1). Такой способ кодирования проецирует категориальные признаки на вещественную прямую. Это приводит к ложным интерпретациям, категориальность данных теряется.

| **region** | **id** |
|---|---|
| Московская область | 1 |
| Самарская область | 2 |
| Пермский край | 3 |

Таблица 1: Кодирование уникальным значением

- Dummy-кодирование [1]. Для категориального признака задаются N новых дихотомических признаков. Получается бинарная матрица, у каждого категориального признака только в одном из дихотомических признаков стоит единица, в остальных нули (см. Таблицу. 2).

| **region** | **region1** | **region2** | **region3** |
|---|---|---|---|
| Московская область | 1 | 0 | 0 |
| Самарская область | 0 | 1 | 0 |
| Пермский край | 0 | 0 | 1 |

Таблица 2: Dummy-кодирование

- Битовое кодирование. Все значения заменяются порядковыми номерами, которые рассматриваются в двоичном виде (см. Таблицу. 3). Недостатком является возникновение ложных связей и интерпретаций.

- Реализация отдельной регрессии для каждой категории. Недостатками подхода являются увеличение числа необходимых регрессий и снижение мощности проверки, поскольку каждая ре-

| region | region1 | region2 |
|---|---|---|
| Московская область | 0 | 1 |
| Самарская область | 1 | 0 |
| Пермский край | 1 | 1 |

Таблица 3: Битовое кодирование

грессия будет реализовываться на меньшей по размеру выборке, чем в случае общего регрессионного уравнения.

В случае с географическими категориями следует рассмотреть еще один способ кодирования. Этот способ дает хорошие результаты, если известно, что значение выходной переменной не просто коррелирует с областью или городом, а изменяется в соответствии с передвижением по карте. К примеру, на аукционах в Москве и области цена обычно падает в 3 раза, в соседних регионах в 2, а в удаленных — на 1,5. Однако, в анализируемом наборе данных такие зависимости установить не удалось, так, например, среднее отношение финальной цены к начальной в Самарской области составляет 0,56, в Пермском крае — 0,57, а в Москве и Московской области — 0,64.

В качестве оптимального способа кодирование был выбран метод dummy-переменных. Именно с его помощью были получены дальнейшие результаты.

## Множественная линейная регрессия

В таблице 4 приведены результаты апробирования модели множественной линейной регрессии на различных входных параметрах.

| **Начальная** цена | **Регион** | **Номер** квартала | **Коэффициент** детерминации |
|---|---|---|---|
| + | − | − | 0,754 |
| + | + | − | 0,883 |
| + | − | + | 0,766 |
| + | + | + | 0,912 |

Таблица 4: Результаты тестирования модели линейной регрессии

Исходя из полученных результатов, можно сделать вывод, что на точность прогноза финальной цены влияют, помимо начальной цены, следующие характеристики лота: «регион» и «номер квартала», причем влияние «региона» существенно сильнее. Модель линейной регрессии именно с этими входными параметрами легла в основу программного продукта.

Программно была реализована возможность просмотра информации по каждому участнику торгов. На странице участника отображается количество его побед и проигрышей в торгах. Прогнозирование финальной цены лота исходя из историй торгов тех людей, которые подали заявки на участие, является весьма точным, коэффициент детерминации такой модели составил 0,92. Но, к сожалению, такая модель не может быть использована на практике, так как списки участников появляются уже после завершения торгов.

## Инструменты

В программном продукте использовались следующие инструменты и технологии:

- Язык программирования C#;
- Фреймворк для машинного обучения Accord.NET [4];
- СУБД MySQL;
- Интерфейс программирования приложений Windows Forms;
- Платформа Deductor Studio для графической визуализации результатов [3].

## Заключение

В статье были рассмотрены, апробированы и оценены различные подходы к прогнозированию финальной цены лота, выставленного на торги на ЭТП, на основе модели линейной регрессии. Изучены методы применения категориальных данных в обучении и тестировании модели линейной регрессии. Были найдены факторы, оказывающих наибольшее влияние на изменение цены лота в ходе торгов. Модель линейной регрессии, использующая эти факторы, стала основой программного продукта, направленного на помощь в участии в торгах.

## Литература


[1] Kaufman, D. Sweet, R. (1974). Contrast coding in least squares regression analysis. American Educational Research Journal, 11, 359–377. http://www.jds-online.com/files/JDS-563.pdf

[2] Norman Matloff University of California, Davis (2017). From Linear Models to Machine Learning Regression and Classification, with R Examples, 43-80. http://heather.cs.ucdavis.edu/draftregclass.pdf

[3] Deductor Studio, Руководство аналитика. http://basegroup.ru/system/files/documentation/guide$_a nalyst_5$.2.0.*pdf*

[4] Accord.NET, Regression. https://github.com/accord-net/framework/wiki/Regression