|  |
| --- |
| Пермский филиал федерального государственного автономного образовательного учреждения высшего образования  «Национальный исследовательский университет  «Высшая школа экономики»  *Факультет экономики, менеджмента и бизнес-информатики* |
|  |
| Холодницкий Сергей Вадимович  Пласкевич Анастасия Викторовна  **ЭКОНОМЕТРИЧЕСКОЕ ИССЛЕДОВАНИИ ДАННЫХ СКРОББЛИНГОВОГО СЕРВИСА LAST.FM**  *Контрольный проект*  по направлению подготовки *38.03.05 Бизнес-информатика*  образовательная программа «Бизнес-информатика»   |  |  | | --- | --- | |  | Руководитель  PhD, к. т. н., доцент кафедры информационных технологий в бизнесе  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  А.В. Бузмаков | |  |  |   Пермь, 2019 год |

**Оглавление**

[Оглавление 2](#_Toc532143161)

[Аннотация 3](#_Toc532143162)

[Введение 4](#_Toc532143163)

[Глава 1. Описание данных 6](#_Toc532143164)

[1.1. Исходные данные 6](#_Toc532143165)

[1.2. Предобработка исходных данных 6](#_Toc532143166)

[1.2.1. Пошаговое описание процесса предобработки набора данных 7](#_Toc532143167)

[1.2.2. Итоговые исходные данные 8](#_Toc532143168)

[1.3. Предварительный анализ данных 8](#_Toc532143169)

[1.3.1. Проверка наличия статистических выбросов 8](#_Toc532143170)

[1.3.2. Проверка однородности данных 11](#_Toc532143171)

[1.3.3. Проверка нормальности распределения переменных 11](#_Toc532143172)

[Эконометрическая модель 16](#_Toc532143173)

[Эмпирические результаты 19](#_Toc532143174)

[Заключение 20](#_Toc532143175)

[Библиографический список 21](#_Toc532143176)

[Приложение 1 . Листинг скрипта для предобработки данных в RStudio 22](#_Toc532143177)

Аннотация

В данной работе представлен результат анализа статистических данных сервиса сбора информации о прослушивании музыки Last.fm по состоянию на 2012 год. Построены линейные регрессионные модели. В конечную выборку вошли данные о прослушиваниях музыки 5 000 пользователей со всего мира.

Работа состоит из введения, трех глав: описание данных, эконометрическая модель и эмпирические результаты, заключения, библиографического списка из 6 наименований и 1 приложения. Объем письменной работы – 22 страницы.

Результат проверки отчета в системе Антиплагиат: 0,64% заимствований, 99,36% - оригинальность, 0% цитирования.

Введение

Сейчас в мире наблюдается тренд сбора данных — тысячи компаний во всех уголках земного шара столкнулись с задачей узнать своего потребителя. Предсказывая поведение потребителя, они хотят повысить свою эффективность, прибыль, выйти на новые рынки и др. Практически каждый сервис имеет внутри специальный механизм для отслеживания поведения пользователя, собирая всевозможные значимые и незначимые факторы о пользователе в особое хранилище. Затем руководством или менеджментом ставятся предположения, некоторые из которых могут быть проверены с помощью инструментов эконометрики, особенно, когда дело касается количественных данных.

В данной работе мы рассмотрим, как минимальная информация о пользователе музыкального сервиса last.fm позволит предсказать его возраст или узнать, от чего зависит приобретение платной подписки на данный сервис. Актуальность вопроса для бизнеса очевидна: увеличение количества платных подписок увеличивает доходы компании и позволяет оптимизировать каналы продаж, настроить таргет-рекламу и другое, чтобы донести необходимую информацию до пользователей с максимальной конверсией в платные подписки.

Таким образом, цель работы – построить линейные по параметрам эконометрические модели для проверки установленных гипотез, путем сравнения моделей по различным показателям — выбрать наилучшую или оптимальную модель для последующей интерпретации новых наборов данных.

Для достижения данной цели были поставлены следующие задачи:

* Проанализировать и обработать исходные данные.
* Построить линейные по параметрам модели согласно гипотезам.
* Проверить модели согласно условиям Гаусса-Маркова, установить мультиколлинеарность и гомо- или гетероскедастичность (Breusch-Pagan test) и выбрать наилучшую или оптимальную модель.
* Проинтерпретировать полученные результаты. Подтвердить или опровергнуть выдвинутые гипотезы.

Выдвигаем следующие гипотезы для их опровержения или подтверждения:

*H0: Возраст отрицательно влияет на объем прослушиваемой музыки.* Идея гипотезы заключается в следующем: традиционно, люди более старшего возраста пользуются меньшим количеством интернет-сервисов, а любовь к музыке принято считать более юным увлечением.

Н1: В*озраст обратно зависит от количества плейлистов и любимых треков пользователя.* Было выдвинуто предположение, что люди более младшего возраста стараются упорядочивать музыку в своих аккаунтах, выбирать и оценивать треки, чтобы сервис им рекомендовал действительно любимую музыку – трепетно относятся к музыке как к части своей жизни.

1. **Описание данных**
   1. Исходные данные

Исследуемый набор данных – данные о был скачан с бесплатного источника Socrata в формате *.csv*.

Тут хочу показать все столбцы мб распределения

* 1. Предобработка исходных данных

Перед тем как анализировать данные, необходимо выполнить их предобработку, для уменьшения размера исходных данных и очищения их от не исследуемых переменных. Предобработка данных была выполнена в программе RStudio. Полный листинг скрипта находится в Приложение 1.

* + 1. Пошаговое описание процесса предобработки набора данных
    2. Итоговые исходные данные
  1. Предварительный анализ данных
     1. Проверка наличия статистических выбросов
     2. Проверка однородности данных

Описательные статистики по всем переменным представлены в разделе *1.3.1*

Создание моделей

* 1. Создание модели

**Эмпирические результаты**

В ходе анализа была выбрана оптимальная модель:

**Заключение**

Библиографический список