**Научно-исследовательский семинар «Анализ данных в Python»**

Контрольная работа №2

Файл **telco.xls** содержит данные о 1000 клиентах телекоммуникационной компании (личная информация и данные об услугах, которыми они пользуются).

|  |  |
| --- | --- |
| **Название переменной** | **Описание переменной** |
| region | Территориальная зона проживания клиента (Zone 1, Zone 2, Zone 3) |
| tenure | Число месяцев использования клиентом сервисов компании |
| age | Возраст клиента в годах |
| marital | Семейное положение клиента (Unmarried, Married) |
| address | Число лет проживания клиента по текущему адресу |
| income | Доход семьи клиента в тысячах у.е. |
| ed | Уровень образования клиента (Did not complete high school, High school degree, Some college, College degree, Post-undergraduate degree) |
| employ | Число лет работы клиента у текущего работодателя |
| retire | Находится ли клиент на пенсии? (No, Yes) |
| gender | Пол клиента (Male, Female) |
| reside | Количество человек в семье клиента |
| tollfree | Toll free service (Yes, No)  Пользуется ли клиент услугой, предполагающей, что на номер можно бесплатно позвонить из любого города страны? |
| equip | Арендует ли клиент оборудование? (Yes, No) |
| callcard | Пользуется ли клиент телефонными картами для звонков? (Yes, No) |
| wireless | Пользуется ли клиент беспроводными сервисами? (Yes, No) |
| longmon | Длительность междугородних звонков в прошлом месяце (в минутах) |
| tollmon | Длительность пользования услугой «Toll free service» в прошлом месяце (в минутах) |
| equipmon | Длительность пользования арендованным оборудованием в прошлом месяце (в часах) |
| cardmon | Длительность пользования телефонной картой в прошлом месяце (в минутах) |
| wiremon | Длительность использования беспроводных сервисов в прошлом месяце (в часах) |
| multline | Multiple lines (Yes, No)  Пользуется ли клиент услугой обслуживания по нескольким линиям? |
| voice | Пользуется ли клиент голосовыми сообщениями (Yes, No) |
| internet | Пользуется ли клиент Интернетом (Yes, No) |
| callid | Пользуется ли клиент услугой определителя номера вызывающего абонента? (Yes, No) |
| callwait | Пользуется ли клиент услугой ожидания звонка (Yes, No) |
| forward | Пользуется ли клиент услугой переадресации звонка (Yes, No) |
| confer | 3-way calling (Yes, No)  Пользуется ли клиент трёхсторонним вызовом |

Файл **sberbank-moscow-exchange.xlsx** содержит ежедневные данные об акциях Сбербанка России.

**! Найдите строку со своей фамилией в нижеприведенной таблице и используйте переменные для анализа в соответствии с рекомендациями.**

|  |  |  |  |
| --- | --- | --- | --- |
| **ФИО студента** | **Бинарная регрессия: зависимая переменная** | **Бинарная регрессия: независимые переменные** | **Анализ временных рядов:**  **переменная и временной период для анализа** |
| Агаркова Полина | tollfree | * region * tenure * age * marital * income   + любые другие 3 переменные | Цена открытия  с 16.07.2007 по 16.07.2010 |
| Альтшулер Эдуард | equip | * marital * age * employ * retire * gender   + любые другие 3 переменные | Максимальная цена  с 16.07.2007 по 16.07.2010 |
| Асатуров Максим | callcard | * age * marital * reside * retire * gender   + любые другие 3 переменные | Минимальная цена  с 16.07.2007 по 16.07.2010 |
| Бакланов Егор | wireless | * income * ed * gender * reside * age   + любые другие 3 переменные | Средняя цена  с 16.07.2007 по 16.07.2010 |
| Балдина Анастасия | multline | * employ * retire * gender * age * marital   + любые другие 3 переменные | Количество сделок  с 16.07.2007 по 16.07.2010 |
| Бердинских Никита | voice | * age * marital * reside * retire * region   + любые другие 3 переменные | Цена закрытия  с 16.07.2007 по 16.07.2010 |
| Бершицкий Дмитрий | internet | * income * ed * gender * marital * age   + любые другие 3 переменные | Цена открытия  с 17.07.2010 по 17.07.2013 |
| Бобышкин Иван | callid | * retire * gender * income * tenure * age   + любые другие 3 переменные | Максимальная цена  с 17.07.2010 по 17.07.2013 |
| Бойко Евгений | callwait | * address * income * region * age * retire   + любые другие 3 переменные | Минимальная цена  с 17.07.2010 по 17.07.2013 |
| Бородулин Арсений | forward | * employ * retire * age * marital * adress   + любые другие 3 переменные | Средняя цена  с 17.07.2010 по 17.07.2013 |
| Бузуруков Саидазамхон | confer | * employ * retire * age * marital * adress   + любые другие 3 переменные | Количество сделок  с 17.07.2010 по 17.07.2013 |
| Городецкий Егор | tollfree | * income * ed * gender * marital * age   + любые другие 3 переменные | Цена закрытия  с 17.07.2010 по 17.07.2013 |
| Кажкаримов Асхат | equip | * income * ed * gender * marital * age   + любые другие 3 переменные | Цена открытия  с 18.07.2013 по 18.07.2016 |
| Кокорев Артём | callcard | * tenure * age * marital * address * employ   + любые другие 3 переменные | Максимальная цена  с 18.07.2013 по 18.07.2016 |
| Кошкин Георгий | wireless | * region * tenure * age * marital * address   + любые другие 3 переменные | Минимальная цена  с 18.07.2013 по 18.07.2016 |
| Лебедев Андрей | multline | * region * gender * tenure * age * marital   + любые другие 3 переменные | Средняя цена  с 18.07.2013 по 18.07.2016 |
| Лукин Илья | voice | * marital * address * income * ed * region   + любые другие 3 переменные | Количество сделок  с 18.07.2013 по 18.07.2016 |
| Можеитов Матвей | internet | * marital * age * region * retire * gender   + любые другие 3 переменные | Цена закрытия  с 18.07.2013 по 18.07.2016 |
| Нгуен Туан | callid | * marital * address * income * ed * region   + любые другие 3 переменные | Цена открытия  с 19.07.2016 по 19.07.2019 |
| Орешкина Наталия | callwait | * retire * gender * reside * region * age   + любые другие 3 переменные | Максимальная цена  с 19.07.2016 по 19.07.2019 |
| Павлов Виктор | forward | * income * ed * reside * marital * age   + любые другие 3 переменные | Минимальная цена  с 19.07.2016 по 19.07.2019 |
| Пасынков Матвей | confer | * marital * address * income * age * region   + любые другие 3 переменные | Средняя цена  с 19.07.2016 по 19.07.2019 |
| Песчанский Дмитрий | tollfree | * ed * employ * tenure * retire * gender   + любые другие 3 переменные | Количество сделок  с 19.07.2016 по 19.07.2019 |
| Петров Константин | equip | * region * tenure * retire * age * gender   + любые другие 3 переменные | Цена закрытия  с 19.07.2016 по 19.07.2019 |
| Полканов Кирилл | tollfree | * age * marital * reside * retire * employed   + любые другие 3 переменные | Цена открытия  с 19.07.2019 по 19.07.2022 |
| Пономарев Николай | equip | * age * marital * reside * retire * employed   + любые другие 3 переменные | Максимальная цена  с 19.07.2019 по 19.07.2022 |
| Сёмкин Арсений | callcard | * age * marital * reside * retire * employed   + любые другие 3 переменные | Минимальная цена  с 19.07.2019 по 19.07.2022 |
| Скляр Наталья | wireless | * age * marital * reside * retire * employed   + любые другие 3 переменные | Средняя цена  с 19.07.2019 по 19.07.2022 |
| Соколов Артём | multline | * age * marital * reside * retire * employed   + любые другие 3 переменные | Количество сделок  с 19.07.2019 по 19.07.2022 |
| Старченков Вадим | voice | * age * marital * reside * retire * employed   + любые другие 3 переменные | Цена закрытия  с 19.07.2019 по 19.07.2022 |
| Тепляков Владислав | internet | * age * marital * reside * retire * employed   + любые другие 3 переменные | Цена открытия  с 15.03.2021 по 15.03.2024 |
| Хадзакос Николай | callid | * age * marital * reside * retire * employed   + любые другие 3 переменные | Максимальная цена  с 15.03.2021 по 15.03.2024 |
| Цимбалов Максим | callwait | * age * marital * reside * retire * employed   + любые другие 3 переменные | Минимальная цена  с 15.03.2021 по 15.03.2024 |
| Чапурина Валерия | forward | * age * marital * reside * retire * employed   + любые другие 3 переменные | Средняя цена  с 15.03.2021 по 15.03.2024 |
| Шикалов Андрей | confer | * age * marital * reside * retire * employed   + любые другие 3 переменные | Количество сделок  с 15.03.2021 по 15.03.2024 |
| Щуплова Анна | callid | * tenure * region * reside * retire * employed   + любые другие 3 переменные | Цена закрытия  с 15.03.2021 по 15.03.2024 |
| Щур Виктория | callwait | * tenure * region * reside * retire * employed   + любые другие 3 переменные | Цена открытия  с 19.07.2020 по 19.07.2023 |
| Анисевич Егор | forward | * tenure * region * reside * retire * employed   + любые другие 3 переменные | Максимальная цена  с 19.07.2020 по 19.07.2023 |
| Пахомов Евгений | confer | * tenure * region * reside * retire * employed   + любые другие 3 переменные | Минимальная цена  с 19.07.2020 по 19.07.2023 |

**Задание 1. Бинарная регрессия**

* Проведите описательный анализ переменных, которые будут использованы в модели бинарной регрессии. Все расчеты должны сопровождаться содержательными комментариями. Нужно провести именно анализ, а не выполнить код.
* Сформулируйте гипотезы исследования, которые будут проверены в ходе проведения регрессионного анализа данных.
* Постройте модель бинарной логистической регрессии. Напишите уравнение регрессии (линейная часть). Оцените статическую значимость и предсказательную силу модели. Постройте ROC-кривую и рассчитайте площадь под ней.
* Оцените значимость регрессионных коэффициентов. Подробно интерпретируйте особенности взаимосвязи между каждым из предикторов и зависимой переменной. Здесь нужно расписать как именно изменение того или иного предиктора изменит шансы положительного исхода.
* Проведите диагностику модели (проверьте остатки на нормальность, выявите выбросы, оцените мультиколлинартность).
* Сформулируйте общие содержательные выводы по результатам регрессионного анализа данных. Какие результаты анализа данных для вас оказались очевидными, а какие неожиданными?

**Задание 2. Анализ временных рядов**

Проанализируйте временной ряд, содержащий в себе ежедневные данные об акциях Сбербанка России. Выполните следующие задания:

* Постройте график, визуализирующий временной ряд. Охарактеризуйте его, присутствуют ли тренд, сезонность или цикл?
* Восстановите пропущенные значения. Обоснуйте и опишите метод, который Вы будете использовать для восстановления пропущенных значений.
* Преобразуйте временной ряд к стационарному виду. Обоснуйте и опишите методы, которые Вы использовали. Приведите график и результаты статистических тестов, подтверждающих, что ряд стал стационарным.
* Постройте и опишите коррелограммы (графики автокорреляционной функции и частной автокорреляционной функции) для временного ряда.
* Разбейте выборку на обучающую и тестовую (обоснуйте выбор временного интервала для тестовой выборки). На основе обучающей выборки идентифицируйте параметры модели АРПСС (ARIMA). Какие параметры модели p, d и q Вы будете использовать и почему? Оцените качество модели. Насколько хорошо она предсказывает значения тестовой выборки?
* Используя модель ARIMA, постройте прогноз временного ряда на 50 периодов вперёд. Как Вы считаете, насколько реалистичный получился прогноз?

**Здание 3 Обучение дата-сайентистов**

Высшее образование в области науки о данных (Data Science) активно развивается в России. Как Вы считаете, какие количественные индикаторы можно использовать для оценки качества образования по науке о данных? Приведите пример исследования, которое позволит выявить факторы, влияющие на качество образования по науке о данных. Сформулируйте цель и гипотезы исследования. Как будет сформирована выборка исследования? Какие переменные и методы анализа данных будут использованы в этом исследовании?

*Обратите внимание, что при выполнении задания нужно высказать свое мнение, опираясь на знания, полученные в процессе изучения дисциплины НИС «Анализ данных в Python». Заимствование теста из сторонних источников и использование больших языковых моделей в этом задании не предполагается.*

**Решения нужно оформить в Jupyter Notebook и загрузить в Smart LMS до 21:00 18 марта. Ноутбук должен содержать выполненный код с результатами расчётов и текстовые блоки с комментариями. Задание выполняется каждым студентом индивидуально.**