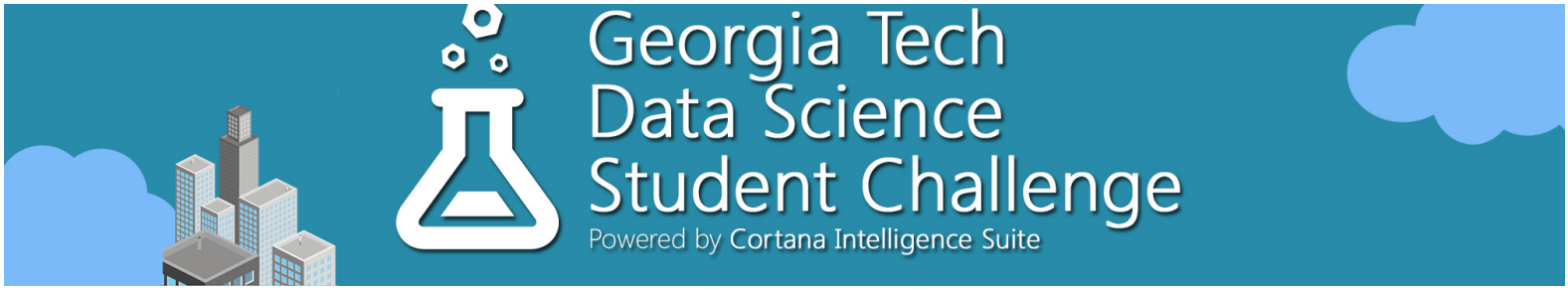


Hackathon Challenge Statement

Theme

Hack & develop an innovative data science solution to help BoomTown gain a competitive advantage on the Real estate market while demonstrating the use of Azure Machine Learning, Microsoft's data science tool.

BoomTown's business model

Founded in 2006, BoomTown is a fast growing, web-based software company which offers a versatile online marketing platform for real estate professionals.

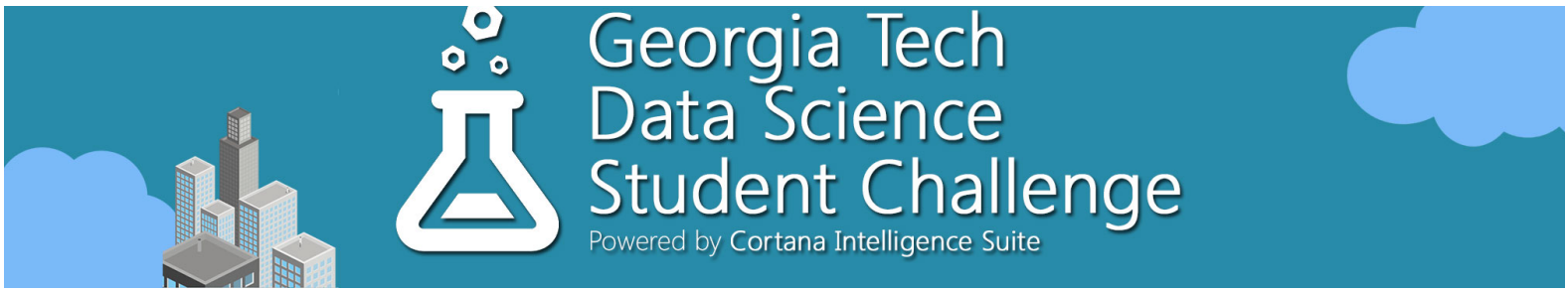
The platform is interfaced with local Multiple Listing Service (MLS) data and includes multiple tools which allow users to automate and personalize sales and marketing efforts continuously.

For instance, the platform contains tools to design and customize brokers' Real Estate websites to attract leads, efficiently interact with sellers and precisely hit the target audience to generate conversions. Others tools are designed to track leads' movements and interactions across the website to identify, predict and act on potential opportunities.

BoomTown's challenges and opportunities

BoomTown's success can be measured by how well they help customers generating conversions and acquiring a competitive advantage in the market.

For instance, BoomTown is interested in leveraging Data & Analytics to:
- Identify potential barriers which prevent properties to be sold : Pricing, Location, Details, Neighbourhood..
- Help brokers in finding the properties they should be dealing: which neighbourhoods, what kind of specific attributes, when should they promote and/or sell it?
- Is the current website design attracting leads and generating conversions?
- What kind of insights about listings, buyers or sellers could brokers leverage to stand out from competition?



Hackathon Instructions

1/ Select one challenge
We came up with 4 distinct challenges to approach the case study.
Teams will be asked to select only one problem among the four problems presented.
Teams have 24 hours to create a solution.
Every team will have 3 minutes to present the proposed solution

For example, one could potentially approach the first problem (Unsold listings) by training a classification model to identify the listing's attributes which have a positive impact on the duration of a (new) listing in the market.
The results of the model could then be used to design the solution that will help BoomTown in creating more value for their customers.

2/ Combine data
Our four challenges can be approached using the dataset that BoomTown has provided. However, hackers are encouraged to blend and enrich that dataset with public datasets and APIs to expend the possibilities and explore new areas that could help BoomTown customers stand out from competition.

3/ Be creative and have fun!

Challenges

Challenge Option 1: Predicting Time to sell for active listings (Offer)

With historically low inventory levels in most parts of the country, understanding how long a listing will remain on market is a critical metric for home buyers to be aware of. There are many opportunities that buyers could miss out on their dream home because they did not act fast enough.

BoomTown would like to predict the remaining days on market for an active listing at any given time and list price in order to help homebuyers understand when they need to make an offer on a home quickly.

Ideas:

- Using BoomTown listing and consumer behavior data, can we predict how long an active listing will remain on the market before it goes under contract?

- What are the attributes of an active listing which have an impact on the number of days it will remain on the market? Is it related to demand, to the attributes of the listing and the location of the listing?

Challenge Option 2: Pricing future Listings (Demand)

To create a great consumer experience, it is also important for real estate professionals to understand how the attributes of a listing will affect the Days On Market (DOM) for a property which has not been listed yet. Displaying and showing popular and relevant listings to buyers can help to sell listings more quickly.

Ideas:

- Given the information about a property which is not actively listed on the market, can we predict how long a listing would be on the market at a given price?
- What are the attributes of new listings which have an impact on DOM ? Is it related to the pricing and location of the listing or maybe the quality of the listing in the website?
- How do these valuable attributes differ by geographic region and listing price ranges?

Challenge Option 3: BoomTown revenue optimization

BoomTown has a large number of listings, some of these might be under or over priced.
This could result in a higher number of unsold listings, missed opportunities and potentially a lower revenue.
Can we use historical pricing data to group properties into detailed micro neighbourhood and show areas with pricing anomalies?
Is it possible to find the optimal pricing based on a listing's attributes?

Ideas

- As an example, can we find the probability of a listing being sold at a specific price?
- Is there a better time to sell a listing and maximize profit?

Challenge Option 4: BoomTown Real estate market knowledge

- BoomTown listings have many attributes that are mostly used for valuation. Most of these attributes are related to the physical characteristics of the listing.
  - But what information about a listing can we gather from a listing relationship to it's street ?
  - For example, is the pricing driven by the pricing of neighbouring listings?
  - Can we find areas with a large number of popular listings that don't necessarily map to standard neighbourhood boundaries?
  - Can we find areas with trends of increasing popularity and absorption rates to identify which territories are most likely to sell?

Datasets

Listing

- Specifies all characteristics associated with a listing. A listing is an individual home, parcel, property, etc. for sale.
- There are a total of 750,000 listing records. 500,000 are sold listings. 250,000 are either active, active contingent, or pending.
- Currently, there over 370 columns on the Listing table, far too many to detail here. The majority of them are bit data types, setting the named column to either true or false. There are also numerous DateTime columns that should be self-explanatory. Columns requiring explanation are detailed below.
- Columns
    - ListingID (bigint)
        - Primary key for the table
    - ListingMLS (varchar (30))
        - The ID provided by the MLS. Non-unique.
    - StatusID (varchar (2))
        - Status of the listing. Possible values:
            - A = Active
            - AC = Active Contingent
            - P = Pending
            - S = Sold
    - CategoryID (int)
        - Property category for the listing. Possible values:
            - 1 = Residential
            - 3 = MultiFamily
    - PropertyTypeID (char (3))
        - Property type of the listing. Possible values:
            - SF = Single Family
            - HRC = High Rise Condo
            - LRC = Low Rise Condo
            - M = Manufactured Home
            - MF = Multi-Family
            - PH = Patio Home
            - VT = Villa/Townhouse
    - ListDate (datetime)
        - Date that the listing was posted to the MLS.
    - PendingDate (datetime)
        - Date that the listing received an accepted purchase offer. ***Use this date for DOM calculations.***

    - SoldDate (datetime)
        - Date that the listing officially closed. Do *not* use for DOM calculations!
- Data mirrors
    - https://georgiatechdata.blob.core.windows.net/georgiatechdata/Listing.csv.gz
    - https://s3.amazonaws.com/boomtown.analytics.dev/listing_extracts/Listing.csv.gz

FeatureList

- Specifies individual listing features.
- Columns
    - ListingID (bigint)
        - Primary key for the table (compound).
        - Foreign key to ListingID in Listing
    - Name (varchar (50))
        - Primary key for the table (compound).
        - Type of listing feature.
    - Value (varchar (500))
        - Description of the listing feature.
    - Topic (varchar (50))
        - Part of the listing that the feature applies to. Possible values:
            - Exterior
            - Exteroir (sic)
            - Hidden
            - Interior
            - Property
- Data mirrors
    - https://georgiatechdata.blob.core.windows.net/georgiatechdata/FeatureList.csv.gz
    - https://s3.amazonaws.com/boomtown.analytics.dev/listing_extracts/FeatureList.csv.gz

VisitorAction

- Specifies all visitor actions taken for a given listing.
- Columns
    - ListingID (bigint)
        - Foreign key to ListingID in Listing.
    - WebVisitorID (bigint)
        - ID of the visitor who performed the action.
    - VisitorAction (varchar (50))
        - Action performed on the listing. Possible values:

        - Asked Question
        - Call From Mobile
        - Emailed
        - Favorited
        - Loan Calculated
        - Printed
        - Schedule Showing
        - Unfavorited
        - Viewed
        - Viewed Again
        - Viewed Facebook Form
        - Viewed PhoneValid
    - ActionDate (datetime)
        - Date the action was taken.
- Data mirrors
    - https://georgiatechdata.blob.core.windows.net/georgiatechdata/VisitorAction.csv.gz
    - https://s3.amazonaws.com/boomtown.analytics.dev/listing_extracts/VisitorAction.csv.gz

ListingChange

- Specifies all price changes for a given listing.
- Columns
    - ListingChangeID (int)
        - Primary key for the table.
    - ListingID (bigint)
        - Foreign key to ListingID in Listing.
    - ChangeDate (datetime)
        - Date that that the listing changed.
    - ChangeType (varchar (50))
        - Type of listing change. Possible values:
            - Listing Sold
            - New Listing
            - Pending
            - Price Increased
            - Price Reduced
    - Description (varchar (1000))
        - Details of the listing change.
- Data mirrors
    - https://georgiatechdata.blob.core.windows.net/georgiatechdata/ListingChange.csv.gz
    - https://s3.amazonaws.com/boomtown.analytics.dev/listing_extracts/ListingChange.csv.gz

Tools

Azure ML: Azure ML is completely free and access only requires signing up to a Microsoft ID (if you don't already have one).

1. Go to https://azure.microsoft.com/en-us/services/machine-learning/
2. Click on Get Started Now
3. Create Free Workspace ($0/Month)

The Hackathon component includes downloading the dataset and uploading into Azure ML. There are a number of options as to how the data can be used in Azure ML, including but not limited to;

- Creating an app
- Further data science analysis and plug into Microsoft Excel
- Plug into Power BI to visualize the solution

Useful Azure ML Resources

Access to Azure Machine Learning and getting Started documentation
https://studio.azureml.net/

Azure Machine Learning Tutorial for data scientists :
https://gallery.cortanaanalytics.com/Experiment/Tutorial-for-Data-Scientists-3

Data science algorithm cheat sheet :
https://azure.microsoft.com/enus/documentation/articles/machine-learning-algorithm-cheat-sheet/

Basic experiment showing use of Azure Machine Learning with Python
https://gallery.cortanaanalytics.com/Experiment/Cortana-Conf-CA-Milk-Python-1?fromlegacydomain=1
https://github.com/Quantia-Analytics/Cortana-Data-Science-Example-Python

Basic experiment showing use of Azure Machine Learning with R
https://azure.microsoft.com/enus/documentation/articles/machine-learning-r-quickstart/

Useful tips on python and r packages for Azure ML
https://microsoft-my.sharepoint.com/personal/akannava_microsoft_com/_layouts/15/WopiFrame.aspx?sourcedoc={2C2E5F49-0904-42DE-A3C6-98A593DDB6D6}&file=Getting%20Started%20with%20Azure%20ML&action=default&d=w2c2e5f49090442dea3c698a593ddb6d6&RootFolder=%2fpersonal%2fakannava%5fmicrosoft%5fcom%2fDocuments%2fShared%20with%20Everyone%2fGetting%20Started%20with%20Azure%20ML

Useful Azure Machine Learning Resources

1. Access to Azure ML and getting Started documentation : https://studio.azureml.net/
2. Azure ML Tutorial for data scientists : https://gallery.cortanaanalytics.com/Experiment/Tutorial-for-Data-Scientists-3
3. Data science algorithm cheat sheet : https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/
4. Basic experiment showing use of AML with Python https://gallery.cortanaanalytics.com/Experiment/Cortana-Conf-CA-Milk-Python-1?fromlegacydomain=1 https://github.com/Quantia-Analytics/Cortana-Data-Science-Example-Python
5. Basic experiment showing use of AML with R https://azure.microsoft.com/en-us/documentation/articles/machine-learning-r-quickstart/
6. Importing data into excel : Import Azure ML data into Excel , Video for PowerQuery in Excel
7. Useful tips on python and r packages for Azure ML https://microsoft-my.sharepoint.com/personal/akannava_microsoft_com/_layouts/15/WopiFrame.aspx?sourcedoc={2C2E5F49-0904-42DE-A3C6-98A593DDB6D6}&file=Getting%20Started%20with%20Azure%20ML&action=default&d=w2c2e5f49090442dea3c698a593ddb6d6&RootFolder=%2fpersonal%2fakannava%5fmicrosoft%5fcom%2fDocuments%2fShared%20with%20Everyone%2fGetting%20Started%20with%20Azure%20ML